

Analysis of Multiple Flows using Different High Speed TCP protocols on a General Network

Sudheer Poojary Vinod Sharma

Department of ECE, Indian Institute of Science, Bangalore

Email: {sudheer, vinod}@ece.iisc.ernet.in

February 23, 2016

Abstract

We develop analytical tools for performance analysis of multiple TCP flows (which could be using TCP CUBIC, TCP Compound, TCP New Reno) passing through a multi-hop network. We first compute average window size for a single TCP connection (using CUBIC or Compound TCP) under random losses. We then consider two techniques to compute steady state throughput for different TCP flows in a multi-hop network. In the first technique, we approximate the queues as M/G/1 queues. In the second technique, we use an optimization program whose solution approximates the steady state throughput of the different flows. Our results match well with ns2 simulations.

Index terms— Internet, High speed TCP protocols, TCP CUBIC, TCP Compound, Multihop network, Performance analysis.

1 Introduction

TCP ensures reliable end-to-end data transfer between hosts and also does flow control and congestion control. While the goal of flow control is to avoid overwhelming the receiver with more packets than it can handle, the role of congestion control is to avoid congestion over the network. A good congestion control scheme must avoid congestion collapse, i.e., overwhelming congestion in the network with severely degraded throughput for all users. It must also be fair and efficient.

Traditional TCP variants (TCP Reno, TCP New Reno) have been very effective in preventing Internet congestion collapse for more than two decades [1]. However as link speeds increase, these TCP variants have been found to be inefficient. The traditional TCP variants use AIMD (additive increase multiplicative decrease) algorithm for congestion window evolution. While AIMD is fair in the sense that all flows going through the same set of links get the same throughput [2], it need not be efficient, especially for high bandwidth-delay product scenarios. The primary reason for the inefficiency being the slow

rate of increase of window size when there is no congestion and the drastic multiplicative drop when there is congestion. Recent years have brought forth high speed variants of TCP (e.g., TCP CUBIC [3], FAST TCP [4], H-TCP [5], Compound TCP [6]). These alter the ‘AI’ phase to make it more aggressive so that links are rarely underutilized. The high-speed TCP variants are now commonly used. TCP CUBIC has been in use on Linux systems since 2006 (Linux kernel 2.6.16) and TCP Compound is used for Windows server. Consequently, measurements in [7] on 30000 web servers reveal that a majority of these web servers ($> 60\%$) use either TCP CUBIC, TCP BIC (the predecessor of CUBIC) or TCP Compound whereas usage of Reno is restricted to less than 15%.

Along with rapid rise in link speeds, we see a rapid increase in the number of mobile devices accessing the Internet. Wireless links are prone to fading, interference and attenuation which may cause random packet losses. Since TCP treats losses as indication of congestion and reduces its window size in response to packet loss, random packet losses considerably deteriorate TCP throughput.

In this paper we develop models for TCP performance in a network with multiple queues and multiple TCP flows (these may use Compound, CUBIC or New Reno) with random packet losses. We will be interested in the performance of long-lived flows (e.g., HTTP streaming, backups, large file downloads). The slow start phase of TCP is of interest when studying performance of short-lived flows but does not impact the performance of long-lived flows if they have low packet rates ($< 1\%$). Since our focus is on performance of the different TCP variants mentioned above, which differ only in the congestion avoidance phase, we will not consider the slow start phase.

1.1 High speed TCP variants: A brief overview

High speed TCP variants alter the congestion avoidance phase behaviour of the traditional TCP to achieve higher efficiency. High speed TCP (HSTCP) [8] addresses the issue of low link utilization in large bandwidth-delay product (BDP) networks by making the window increments and decrements a function of the current window size. Scalable TCP [9] replaces additive increase by multiplicative increase so that the TCP flow rate converges quickly to the link speeds in large BDP networks. H-TCP [5] alters the window increments so that increments depend on time since last congestion. BIC congestion control [10] uses binary search to obtain an efficient operating point. FAST TCP [11] is different from the earlier described variants in the sense that it uses queuing delay as a measure of congestion unlike the other variants which use packet loss. It can be considered to be a high speed variant of TCP Vegas[12]. FAST TCP uses variable size increments and decrements based on the estimate of queuing in the network. In [3], the authors propose TCP CUBIC congestion control algorithm. Like H-TCP, TCP CUBIC window size is a function of time elapsed since last congestion. The authors choose a cubic function for window evolution. TCP Compound [6] is a delay-based congestion control algorithm. Delay-based congestion control algorithms experience lower losses, lower queuing delays and better RTT-fairness as compared to loss-based congestion control algorithms

[11], [13]. However, in a mixed environment, when competing with flows using loss-based congestion control, they are not able to get their fair share of capacity. To address this drawback, TCP compound window also has a loss-based component which gives it a worst-case performance of TCP New-Reno. An extensive survey describing the different TCP variants can be found in [14].

There is a considerable amount of literature on simulation and experimental evaluation of the high speed TCP variants. In [15], the authors perform experimental evaluation of TCP CUBIC in a small buffer regime. Results in [16] show that the ns2 implementation results for several high speed TCP variants match well with experimental results. The intra-protocol and inter-protocol fairness of different high speed TCP variants has been studied in [17] and [18] using ns2 simulations. The Reference [19] is a recent experimental survey of high speed TCP fairness. It evaluates fairness of TCP SACK, HSTCP and CUBIC TCP on a 10 Gbps optical link. They show that fairness is a function of buffer sizes and queue management schemes at the routers with RED routers yielding more equitable rate allocations than droptail.

1.2 Analytical studies of TCP

Traditional TCP (TCP Tahoe, TCP Reno and TCP New Reno) has been extensively studied and analyzed. In [20], the authors use a periodic loss model to compute the average window size of TCP as a function of the packet error probability. The reference [21] uses Markov regenerative processes to model the window evolution of TCP Reno (congestion avoidance phase) under random losses. In [22], the authors consider the effect of connection establishment, slow start and congestion avoidance phase on TCP latency. In [23], the authors compute the throughput and mean sojourn time of TCP Tahoe and TCP Reno flows over a single bottleneck link using RED queue management. The link also carries UDP traffic which has priority over the TCP traffic. In [24], the authors prove stability of multiple TCP Tahoe and TCP Reno flows passing through a single drop-tail or RED queue in the presence of UDP traffic. They then extend their results to the situation when there are multiple bottleneck links in tandem. In [25], the authors use Markovian models to compute mean download times for ON-OFF TCP Tahoe and TCP Reno flows and throughput for long-lived TCP flows in the presence of UDP traffic.

An optimization-based approach is used to analyze network congestion control in [26, 27, 28, 29, 30]. In [26], the authors consider a generic network-wide global optimization problem and derive a distributed congestion control algorithm whose equilibrium rate allocations are a solution to the global optimization problem. As opposed to this approach, [27] and [28] start with distributed congestion control algorithms and derive the corresponding network-wide global optimization problem. In [29], under a AIMD TCP-like scheme with fixed window size, the authors show that FIFO queuing gives proportional-fair rate allocation, longest queue first gives maximum sum-rate and fair queuing policy yields max-min fair rate allocation. An analytical model for identifying the bottleneck links in a multi-hop network is given in [30].

Differential equations are used to model TCP behaviour in [31, 32, 33, 34]. In [31], the authors compute steady state throughput of TCP Tahoe and TCP Reno under random losses. In [24] and [32], the authors model transient behaviour of RED routers supporting TCP flows. A mean-field model for TCP is developed in [33] for multiple TCP flows sharing a single bottleneck link which employs RED queue management. Under the same setup of single bottleneck link and multiple TCP flows, the authors in [34] show that in the many flows regime the data rate evolution and drop rate evolution at the queues can be approximated by a deterministic system.

1.3 Previous studies of TCP CUBIC and Compound

The newer variants of TCP have fewer analytical studies. In [35], the authors use a Markovian model to compute steady state throughput of a single TCP CUBIC connection in a wireless environment. A mean field model is used for performance analysis of multiple TCP CUBIC connections going through a single drop-tail bottleneck link in [36]. In [37], the authors compute throughput of a single long-lived Compound TCP under random losses through a Markovian model. There are deterministic models for computation of average window size of CUBIC and Compound TCP in [3] and [6] respectively. One advantage that these models have over the Markovian models is that they provide a closed-form expression for the average window size of a TCP flow in terms of its RTT and packet error rate. We have investigated these closed form expressions against our Markovian models in [38] and have found that the closed form expressions are not as accurate as the Markovian model results. In [39, 40, 41], the authors study the performance of TCP Compound using control theoretic techniques and derive stability conditions for TCP Compound. In these papers, it is shown that when multiple TCP flows share a single bottleneck queue, the queue sizes and the link utilization have oscillatory behaviour when the feedback delays (round trip times of the flows) and buffer sizes are large. In [39], the authors evaluate the TCP Compound performance as a function of buffer size in the bottleneck queue. In [40], the authors study the performance of TCP Compound with a proportional integral enhanced queue management policy whereas in [41], RED queue management policy is considered.

1.4 Our Contribution

Markov models for TCP CUBIC and TCP Compound do exist in [35] and [37] respectively. However, the Markov model for TCP CUBIC in [35] assumes a different loss model; the inter packet loss durations are assumed Poisson. In our setup, we assume that packets are lost independently of other packets. This scenario is close to the approach used in [21, 26]. We note that the Markov model for TCP Compound in [37] is similar to our model. However, our Markov model for TCP Compound, unlike in [37], also includes the queue lengths resulting in better approximation in real world scenario. Also, our Markov chain models

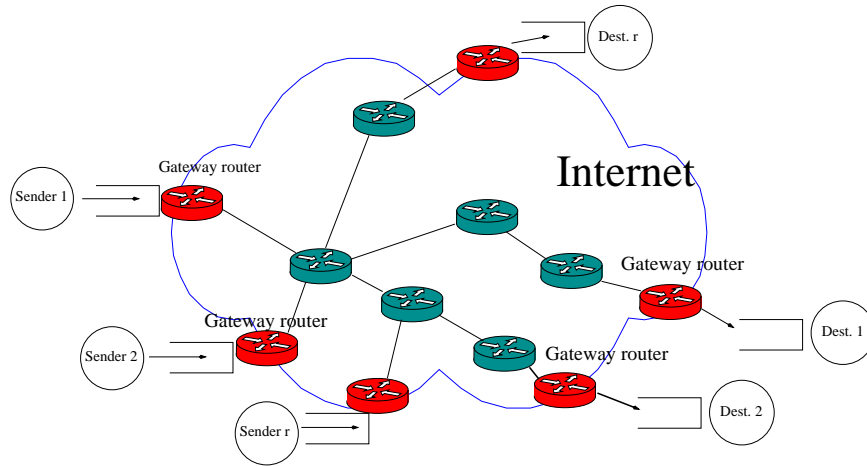


Figure 1: The general network, $(\mathcal{R}, \mathcal{L})$

for TCP CUBIC and TCP Compound have less computational complexity than those in [35] and [37].

In this paper, we develop techniques for performance analysis of different TCP flows (using TCP Compound, TCP CUBIC or Reno) over a general network with multiple bottleneck links. There are a number of simulation and experimental evaluations for performance analysis of these TCP variants over different network topologies. However, to the best of our knowledge, our work is the first theoretical model for the joint performance analysis of these high speed TCP variants over a general network with multiple bottleneck links. Since presence of different types of TCPs affects the throughputs of each other in complicated ways and this is the practical scenario in the current Internet, it is important to study this setup. We validate our model approximations via extensive ns2 simulations.

The organization of our paper is as follows. In Section 2, we describe our system model. We describe a Markovian model to compute average window size of a single TCP CUBIC connection with fixed RTT under random losses in Section 3. In Section 4, we develop a model for computing the average window size of a single TCP Compound connection with non-negligible queuing under random losses. In Section 6, we describe two techniques viz., M/G/1 approximation and an optimization based approach to compute the steady state average window size and the throughputs for TCP flows (which could be TCP CUBIC, TCP Compound or TCP New Reno) over a multi-hop network. We compare our theoretical results with simulations in Section 7. Section 8 concludes our paper.

2 System Model

Consider a general network of routers as shown in Figure 1. A set \mathcal{R} of TCP flows is passing through this network. The TCP flows are carrying long files. A TCP connection may be using TCP New Reno, TCP Compound or TCP CUBIC. We denote the set of links by \mathcal{L} . Some of the access links may be wireless. For a flow $r \in \mathcal{R}$, let Δ_r be the constant round-trip delay (this includes propagation and transmission delays at the links). Let each packet of flow r be lost with probability p_r on its path, independently of others. This indicates that the packet losses in our system are mainly due to transmission errors on the wireless links and neglects the buffer overflows. This is increasingly the scenario in practical networks. We also assume that the packet losses for different flows are independent.

Let A be the incidence matrix for the network with $A(l, r) = 1$ if packets of flow r go through link l . It is possible that besides the TCP packets, the TCP ACKs are also subject to queuing delays (due to congestion on the path from destination to source). Let B be the incidence matrix for the TCP ACK packets, i.e., $B(l, r) = 1$ if TCP ACK packets of flow r go through link l .

We would like to compute the throughput of each TCP connection in this setup. For this, first, in Sections 3 and 4, we derive average window size for a single TCP flow using TCP CUBIC and TCP Compound, respectively, through a single bottleneck link. Then, using these, in Section 6, we describe techniques to compute the steady state throughputs attained by the different TCP flows in the general network.

3 A Markov Model for TCP CUBIC

In this section, we develop a model for a single TCP CUBIC connection with constant round trip time (RTT). Any packet received can be in error with probability p independent of other packets. This is a realistic assumption for wireless links and is commonly made for TCP models in the literature ([20], [21], [22], [31], [35], [37],). We assume that only TCP data packets are lost. The ACKs are typically smaller in size and hence are less likely to be in error. However, the model can be easily extended to the case where ACKs may also be subject to random losses.

We are interested in the performance of long-lived flows (such as file transfers and video streaming) as these are quite common over the current Internet. Hence, we will ignore the slow start phase of window evolution which does not impact throughput for long-lived flows. Also, TCP CUBIC and TCP Compound (also other high speed TCP variants) only differ in the congestion avoidance phase.

In the congestion avoidance phase, TCP CUBIC uses a non-linear cubic function, (W_{cubic} in equation (1)) for window evolution. However the window sizes obtained by the cubic function can at times be smaller than TCP Reno. In such a scenario, TCP CUBIC uses another function, $W_{reno}(t)$ for window

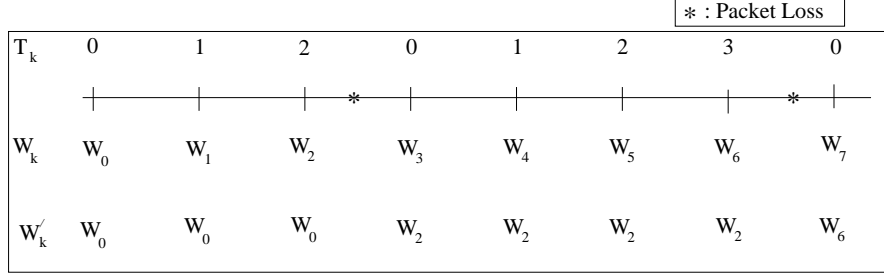


Figure 2: TCP CUBIC: Illustrating $\{W_n, W'_n, T_n\}$ processes

evolution. Assuming that at time $t = 0$, there is a packet loss and there are no further losses in $(0, t)$, the window size, $W(t)$ of TCP CUBIC at time t is given by $W(t) = \max\{W_{cubic}(t), W_{reno}(t)\}$, where

$$\begin{aligned}
 W_{cubic}(t) &= C \left(t - \sqrt[3]{\frac{W_0 \beta}{C}} \right)^3 + W_0, \\
 W_{reno}(t) &= W_0(1 - \beta) + 3 \frac{\beta}{2 - \beta} \frac{t}{R},
 \end{aligned} \tag{1}$$

and W_0 is the window size at time $t = 0$, C is a constant, β is the multiplicative drop factor and R is the round trip time of the connection. If there is a packet loss at time t , the window size is reduced to $(1 - \beta)W(t)$.

Although in principle the equations in (1) are evolving continuously in time, because of constant RTT, we can assume that the TCP source updates its window size at the beginning of each RTT and transmits that many number of packets. This is usually assumed in TCP studies and will be validated via simulations. The evolution of TCP CUBIC, as given by (1) can be modeled using a simple Markov chain. Let $W_n \in \{1, 2, \dots\}$ denote the window size at the beginning of the $(n + 1)^{st}$ RTT. Let W'_n denote the window size at the last packet loss epoch prior to the end of the n^{th} RTT and T_n denote the number of RTT epochs elapsed between the last packet loss and the n^{th} RTT. If there is no loss between the n^{th} and the $(n + 1)^{st}$ RTT, $W'_{n+1} = W'_n$ and $T_{n+1} = T_n + 1$. If there is a loss we use (1) and set

$$W_n = \max \left\{ C \left(RT_n - \sqrt[3]{\frac{W'_n \beta}{C}} \right)^3 + W'_n, W'_n(1 - \beta) + 3 \frac{\beta}{2 - \beta} T_n \right\}, \tag{2}$$

$W'_{n+1} = W_n$ and $T_{n+1} = 0$. The probability of no loss between the n^{th} and the $(n + 1)^{st}$ RTT is given by $(1 - p)^{W_n}$. We illustrate the $\{W_n, W'_n, T_n\}$ processes in Figure 2. In the Figure, we assume that there is a packet loss just before the first RTT. On packet loss, the value of W'_n is updated and T_n is set to 0. For example, since there is a loss in RTT 3 in Figure 2, at the end of the 3^{rd} RTT, W'_3 is set to W_2 and T_3 is set to 0. Also, W_3 is set to $W_2(1 - \beta)$.

The TCP window size is usually restricted by the buffer size available at the receiver. Let the window size $W_n \leq W_{max} < \infty$. Therefore $W'_n \leq W_{max}$. Let T_{max} be the maximum time (in multiples of RTT) taken for the TCP CUBIC window size, $W(t)$ to hit W_{max} starting from any initial window size $W_0 \in \{1, 2, \dots, W_{max}\}$. We note that if T_n exceeds T_{max} , then W_n as computed by (2) exceeds W_{max} and in this case we set W_n to W_{max} . Therefore, we can restrict T_n to be in $\{1, 2, \dots, T_{max}\}$. Thus the process $\{W'_n, T_n\}$ forms a finite state discrete time Markov chain. From any state in the state space, a sequence of consecutive packet drops would cause the Markov chain to hit $(1, 0)$. Therefore, the state $(1, 0)$ can be reached with positive probability from any state in the state space and hence is recurrent. All states that can be reached from $(1, 0)$ are recurrent and the remaining states are transient. Also, there is a self loop from state $(1, 0)$ to itself. Therefore the process $\{W'_n, T_n\}$ is aperiodic with a single positive recurrent communicating class. Hence it has a unique stationary distribution which we denote by $\pi(w, d)$. Also, starting from any initial state the chain converges exponentially to the stationary distribution in total variation.

Let $\mathbb{E}[W]$ denote the mean window size under stationarity. For TCP CUBIC, it can be computed as

$$\mathbb{E}[W] = \sum_{1 \leq w \leq W_{max}, d \in \mathcal{D}} W(w, d) \pi(w, d), \quad (3)$$

where $\mathcal{D} \subset \{0, 1, \dots, T_{max}\}$ and $W(w, d)$ is given by

$$W(w, d) = \max\left\{C\left(Rd - \sqrt[3]{\frac{w\beta}{C}}\right)^3 + w, w(1 - \beta) + 3\frac{\beta}{2 - \beta}d\right\}. \quad (4)$$

From (2), we see that the transitions of the process $\{W'_n, T_n\}$ (both the next state and the transition probabilities) depend on the RTT of the connection. Therefore the stationary distribution, $\pi(w, d)$ and the mean window size of TCP CUBIC is a function of the RTT of the connection. We denote the relationship between RTT, R , packet error rate, p and the mean window size, $\mathbb{E}[W]$ for TCP CUBIC as

$$\mathbb{E}[W] = f_p(R), \quad (5)$$

which is numerically computed using (3) (see Figures 6 and 7 below). We note that (5) is not a closed form expression but a numerical evaluation of (3). We use (5) in Section 6 where we study TCP CUBIC in a multi-hop network where the RTT may not be constant due to queuing in the network. In that case, we use an approximation and replace R by the mean RTT, $\mathbb{E}[R]$, of the flow.

We do not consider the case of non-negligible queuing for TCP CUBIC. However, in Section 6, where we consider TCP connections with non-negligible queuing, we approximate the TCP CUBIC mean window size $\mathbb{E}[W] \approx f_p(\mathbb{E}[R])$, where $\mathbb{E}[R]$ is the mean RTT of the connection.

3.1 Simulation Results

In Figures 3, 4 and 5, we plot the mean window size $\mathbb{E}[W]$ as a function of W_{max} and compare it to ns2 simulations. In our simulations and model, we

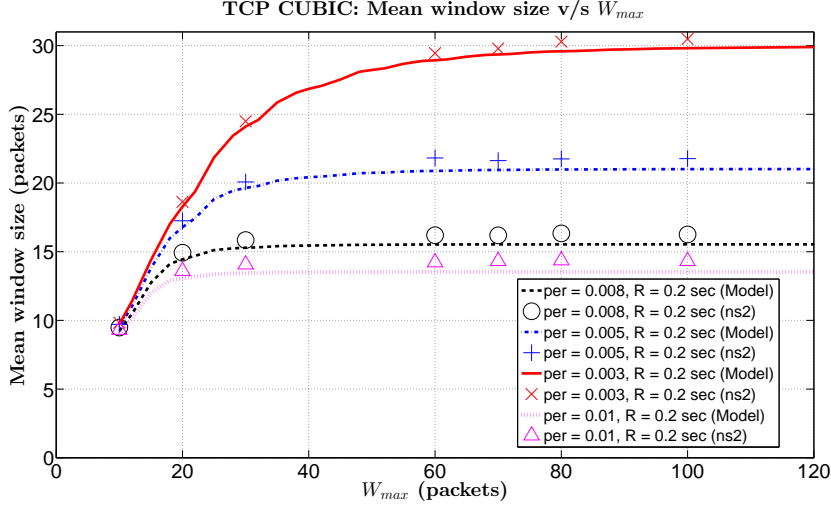


Figure 3: TCP CUBIC: Effect of W_{max} on $\mathbb{E}[W]$.

set $C = 0.4, \beta = 0.3$, which are the values used by the current version of TCP CUBIC [42]. The packet sizes are set to 1050 bytes which is the default value in ns2. The link speeds are set to 1 Gbps. Each packet can be dropped independently of other packets with probability p . The loss in the network is modelled as random. Hence, in all the simulations, we set the link buffer size to be greater than W_{max} so that there are no buffer drops. In Figure 3, we plot results for packet error rates, 0.01, 0.008, 0.005 and 0.003 and RTT is set to 0.2 sec (bandwidth delay product = 23810 packets). In Figure 4, we plot results for packet error rates, 0.001, 0.0003 and 0.0001 and RTT is set to 0.02 sec (bandwidth delay product = 2381 packets). In Figure 5, we plot results for packet error rates, 5×10^{-5} , 3×10^{-5} and 1×10^{-5} and RTT is set to 0.2 sec (bandwidth delay product = 23810 packets). The theoretical and ns2 results differ by $< 8.5\%$. The mean window size $\mathbb{E}[W]$ increases monotonically with W_{max} . However for large values of W_{max} , change in W_{max} has negligible effect on $\mathbb{E}[W]$. This suggests that for large values of W_{max} , $\mathbb{E}[W]$ is not affected by W_{max} . In the rest of the section, we will be working in the regime where W_{max} has no effect on $\mathbb{E}[W]$.

In Figures 6 and 7, we plot our results for mean window size for TCP CUBIC as a function of the round trip time (RTT) and compare them to ns2 simulations. The packet sizes are set to 1050 bytes. The link speeds are set to 1 Gbps. The bandwidth-delay product in these simulations range from 1190 to 59523 packets. The simulation and theoretical results differ by $< 8\%$. We see in these figures that, unlike TCP Reno, the mean window size of TCP CUBIC is a function of the RTT along with the packet error rate. Traditional TCP viz., AIMD TCP (e.g., TCP Reno, TCP New Reno) is known to be unfair to TCP connections with longer RTT. Since the mean window size of CUBIC grows with RTT, TCP

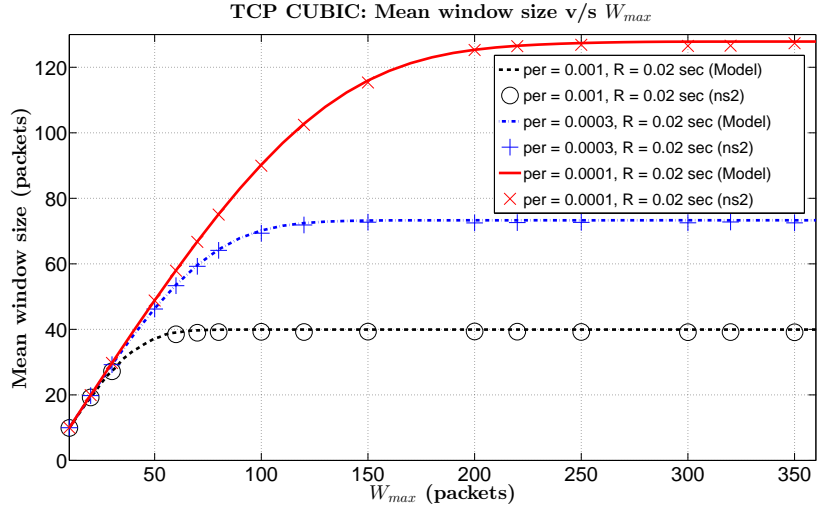


Figure 4: TCP CUBIC: Effect of W_{max} on $\mathbb{E}[W]$.

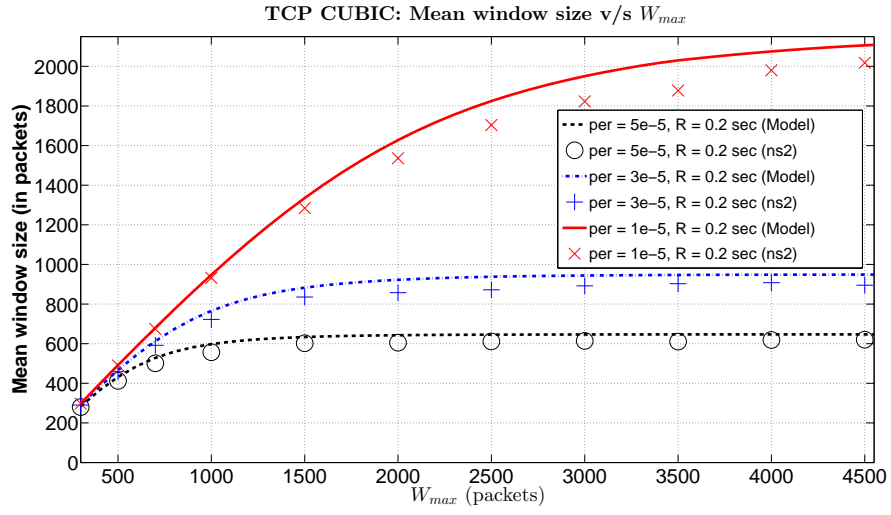


Figure 5: TCP CUBIC: Effect of W_{max} on $\mathbb{E}[W]$.

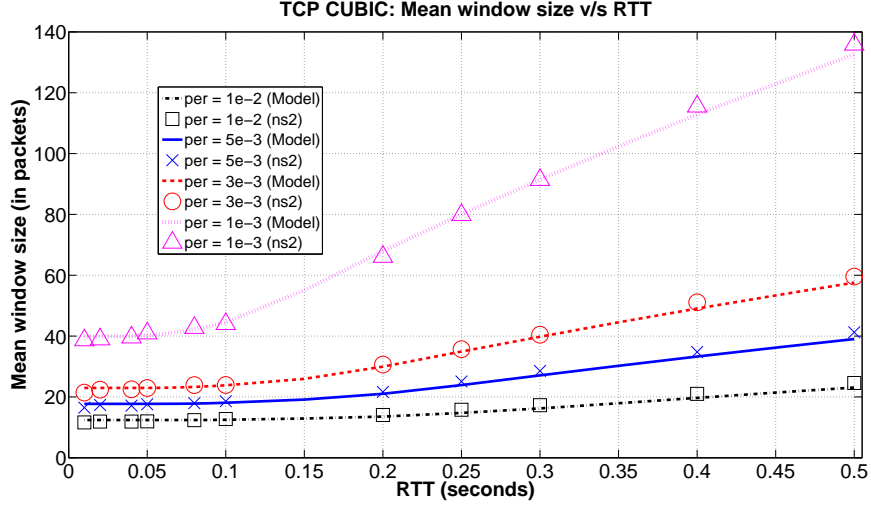


Figure 6: TCP CUBIC: Effect of RTT on $\mathbb{E}[W]$.

CUBIC has better RTT-fairness than AIMD TCP. This feature also makes TCP CUBIC efficient in networks with larger RTT.

4 A Markov Model for TCP Compound

In this section, we develop Markov models for TCP Compound connections with negligible and non-negligible queuing. As before, we have a single TCP connection. We assume that any packet received can be in error with probability p independent of other packets. Also, we assume that ACKs are not lost.

TCP Compound is a delay-based congestion control algorithm. When competing with flows using loss-based congestion control algorithms, flows using delay based congestion control algorithms get less than their fair share of capacity. To address this drawback, TCP Compound window has a loss based component besides the delay based component. The delay based component increases rapidly when there is no congestion in the network and decreases when there is congestion. The loss-based component ensures that when there is congestion in the network, TCP Compound flows behave like TCP-Reno, thus getting their fair share of the network capacity. We denote the window size of TCP Compound at the end of the n^{th} RTT by W_n and the delay based and loss based components by D_n and L_n respectively. The window evolution of TCP Compound is given by

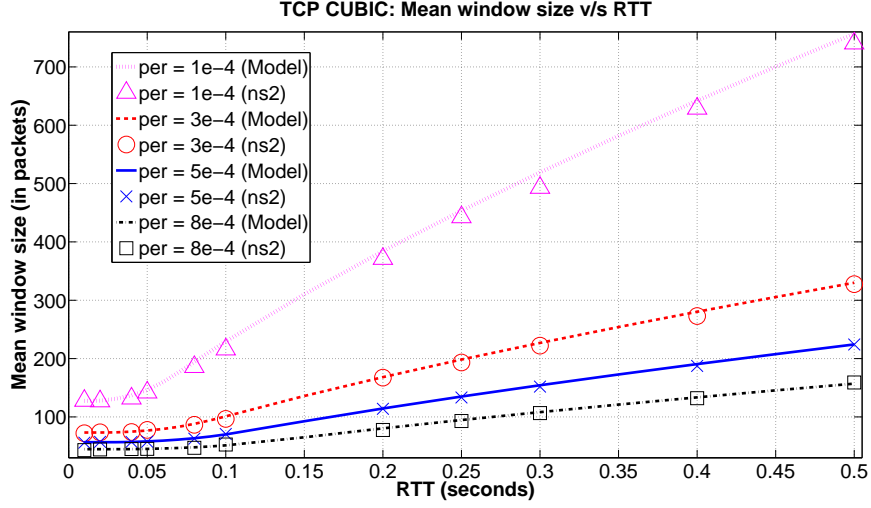


Figure 7: TCP CUBIC: Effect of RTT on $\mathbb{E}[W]$.

$$D_{n+1} = \begin{cases} D_n + (\alpha(W_n)^k - 1)^+, & \text{if no loss during the RTT and } Q_{n+1} < \gamma; \\ (D_n - \zeta Q_{n+1})^+, & \text{if no loss during the RTT and } Q_{n+1} \geq \gamma; \\ \frac{D_n}{2}, & \text{if loss is detected;} \end{cases} \quad (6)$$

$$L_{n+1} = \begin{cases} L_n + 1, & \text{if no loss;} \\ \frac{L_n}{2}, & \text{if a loss is detected;} \end{cases} \quad (7)$$

$$W_{n+1} = D_{n+1} + L_{n+1}; \quad (8)$$

where α and k are constant parameters and γ is a queuing threshold. We see that the loss-based component L_n has behaviour similar to TCP Reno. The delay based component D_n increases aggressively when there is no queuing but decreases when queuing increases beyond threshold γ . The variable Q_{n+1} is the estimate of queuing in the network at the end of the $(n+1)^{st}$ RTT.

4.1 Markov Model for Negligible Queuing

When there is no queuing, the window evolution given in equations (6), (7) and (8) simplifies to

$$W_{n+1} = \begin{cases} W_n + 1 + (\alpha W_n^k - 1)^+, & \text{if there is no loss;} \\ \frac{W_n}{2}, & \text{if there is loss.} \end{cases} \quad (9)$$

The probability of no loss between the n^{th} and the $(n+1)^{st}$ RTT is given by $(1-p)^{W_n}$, whereas the probability of loss is given by $1-(1-p)^{W_n}$. Thus, $\{W_n\}$ is a finite state discrete time Markov chain with $W_n \leq W_{max}$.

The state 1 can be reached with positive probability after a sequence of consecutive drops from any state in the state space and hence is recurrent. All states that can be reached from 1 are recurrent and the remaining states are transient. Also, the Markov chain is aperiodic as the state 1 has a self-loop with self loop transition probability p . Hence, the Markov chain has a unique stationary distribution which can be used to compute the average window size, $\mathbb{E}[W]$. Also, starting from any initial state the chain converges exponentially to the stationary distribution in total variation.

From (9), we see that, when queuing is negligible, unlike TCP CUBIC, for TCP Compound the next state and the transition probabilities are independent of the RTT, R of the connection. Therefore, in this case, the TCP Compound average window size is independent of the RTT of the connection.

4.2 Markov Model for Non-negligible Queuing

We now consider the case when the queuing is not necessarily negligible. In this case, we have to consider both the delay and the loss based components of the window size W_n . Given D_n , L_n and Q_{n+1} , the components D_{n+1} and L_{n+1} can be computed from equations (6) and (7) respectively. We approximate the queue size, Q_{n+1} at the end of the $(n+1)^{st}$ RTT by $Q_{n+1} \approx (W_n - \mu\Delta)^+$ where μ is the bottleneck link capacity in packets/sec and Δ is the constant propagation delay. This approximation is based on the assumption that, when the window size is W , the instantaneous rate at which packets are sent is $\min\{\frac{W}{\Delta}, \mu\}$. Thus the queue size is $(W - \min\{\frac{W}{\Delta}, \mu\}\Delta) = (W - \mu\Delta)^+$. We validate this approximation by simulation results. The process $\{(D_n, L_n)\}$ forms a Markov chain. The transitions are given by equations (6) and (7). The probability of no loss between the n^{th} and the $(n+1)^{st}$ RTT is given by $(1-p)^{W_n}$, whereas the probability of loss is given by $1-(1-p)^{W_n}$. We assume that the window size is upper bounded by W_{max} . Thus the Markov chain has finite state space.

From any state in the state space, a sequence of consecutive packet drops would cause the Markov chain to hit $(0,1)$. Therefore, the state $(0,1)$ can be reached from any state in the state space with positive probability. Hence $(0,1)$ is positive recurrent and all states that can be reached from $(0,1)$ are positive recurrent and the remaining states are transient. Also the state $(0,1)$ has a self loop with self-loop transition probability p . Therefore the Markov chain $\{(D_n, L_n)\}$ is aperiodic and has a unique stationary distribution which we denote by $\pi(d, l)$. Also, starting from any initial state the chain converges exponentially to the stationary distribution in total variation.

Let us denote the stationary value of the $\{W_n\}$ process, where $W_n = D_n + L_n$, by \bar{W} . Suppose R_n denotes the RTT of the packets in the window transmitted at the end of n^{th} RTT. We can approximate R_n as $R_n \approx \max\{\Delta, \frac{W_n}{\mu}\}$. This approximation is quite commonly made [13], [37]. Let \bar{R} be a random variable

with the stationary distribution of $\{R_n\}$. Let us denote the stationary window size for the continuous time window size process, $W(t)$ by \mathcal{W} . Let \hat{R}_k be the RTT for the k^{th} packet. We denote the stationary value for \hat{R}_k by \hat{R} . The time average window size and the packet average RTT for the connection can be computed using Palm calculus [43] as

$$\mathbb{E}[\mathcal{W}] = \frac{\mathbb{E}[\overline{W\overline{R}}]}{\mathbb{E}[\overline{R}]}, \quad \mathbb{E}[\hat{R}] = \frac{\mathbb{E}[\overline{W\overline{R}}]}{\mathbb{E}[\overline{W}]}, \quad (10)$$

where the averages $\mathbb{E}[\overline{W\overline{R}}]$, $\mathbb{E}[\overline{R}]$ and $\mathbb{E}[\overline{W}]$ are computed using $\pi(d, l)$.

4.3 Simulation Results

The results obtained using the Markov model for Compound TCP with negligible queuing are compared with ns2 simulations in Figures 8, 9 and 10. The packet sizes are set to 1050 bytes. The link speeds are set to 1 Gbps so that there is negligible queuing. Each packet can be dropped independently of other packets with probability p . We model the losses in the network as random. Hence, in all the simulations, we set the link buffer size to be greater than W_{max} , so that there are no buffer drops. In Figure 8, we plot results for packet error rates, 0.008, 0.005 and 0.003. In Figure 9, we plot results for packet error rates, 0.001, 0.0008 and 0.0003. In Figure 10, we plot results for packet error rates, 5×10^{-5} , 3×10^{-5} and 1×10^{-5} . We use two different values for the propagation delay, viz., 0.2 sec (bandwidth delay product = 23810 packets) and 0.02 sec (bandwidth delay product = 2381 packets) for Figures 8 and 9 and 0.2 sec (bandwidth delay product = 23810 packets) and 0.1 sec (bandwidth delay product = 1190 packets) for Figure 10. In our model, the average window size does not depend on the RTT of the flow. In Figures 8, 9 and 10 we observe that for the ns2 simulations, the average window size of TCP Compound flows with same packet error rate but different RTT are close to each other. Thus, we see that when the RTT is constant, i.e., queuing is negligible, the average window size of TCP Compound flow does not depend on its RTT. Also, as in the case of TCP CUBIC, for large values of W_{max} , there is no change in the average window size as W_{max} changes. In the rest of the section, we will be working in this regime, i.e., we choose W_{max} large enough so that there is no effect of W_{max} on the average window size.

In Figures 11 and 13, we plot the mean window size for TCP Compound with bottleneck link speed $C = 1$ Mbps. In Figures 12 and 14, we plot the normalized link utilization in this case. The packet sizes are set to 1050 bytes. The bandwidth delay product in these simulations range from 1 packet to 60 packets. In Figure 15 and 16, we plot the mean window size and normalized link utilization with bottleneck link speed $C = 10$ Mbps. The bandwidth delay product in these simulations range from 12 packet to 595 packets. The simulation and model results differ by $< 10.5\%$.

In Figures 13 and 15, we see that as the round trip propagation delay, Δ increases, the average window size of TCP Compound increases. However the

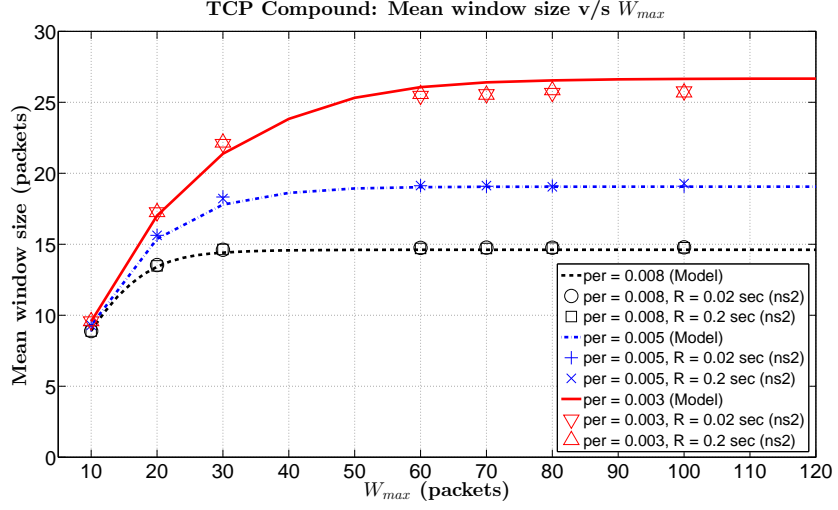


Figure 8: TCP Compound: Effect of W_{max} on $\mathbb{E}[W]$, negligible queuing.

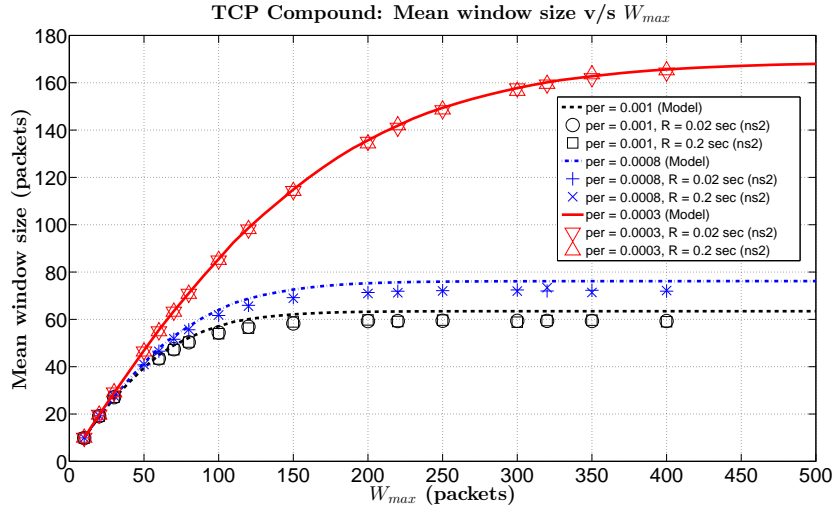


Figure 9: TCP Compound: Effect of W_{max} on $\mathbb{E}[W]$, negligible queuing.

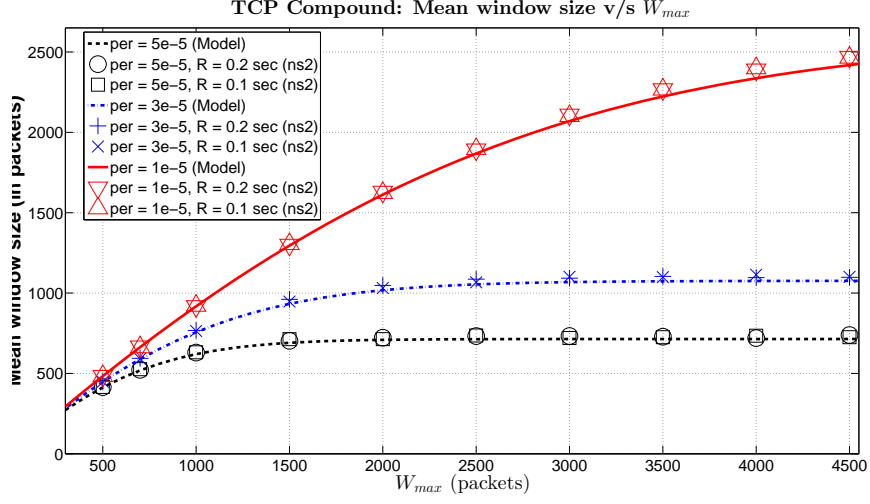


Figure 10: TCP Compound: Effect of W_{max} on $\mathbb{E}[W]$, negligible queuing.

change in average window size is not due to change in Δ , but due to queuing. Due to the delay based component of Compound TCP, flows which encounter more queuing have smaller average window sizes. For a fixed bottleneck link capacity, flows with larger propagation delays have smaller throughput and hence smaller queues at the bottleneck link (by Little's law). Therefore when there is non-negligible queuing, flows with larger propagation delays have larger average window sizes. When window sizes are small, the delay based component has negligible contribution to the window size and TCP Compound behaves like Reno. This behaviour can be seen in Figure 11 for packet error rates of 0.01, 0.008 and 0.005 where there is not much change in mean window size as Δ changes. In Figures 12, 14 and 16, we see that as propagation delay, Δ increases, link utilization decreases. For larger Δ , the random packet errors become a bottleneck and adversely affect link utilization.

The average window size for TCP Compound is a function of packet error rate and the queuing that the flow causes in the network. We use the following approximation for TCP Compound average window size

$$\mathbb{E}[W] \approx g_p(\mathbb{E}[Q]), \quad (11)$$

where p is the packet error rate for the flow and $\mathbb{E}[Q]$ is the average number of packets in the queue. The equation (11) is not a closed form expression but is obtained numerically.

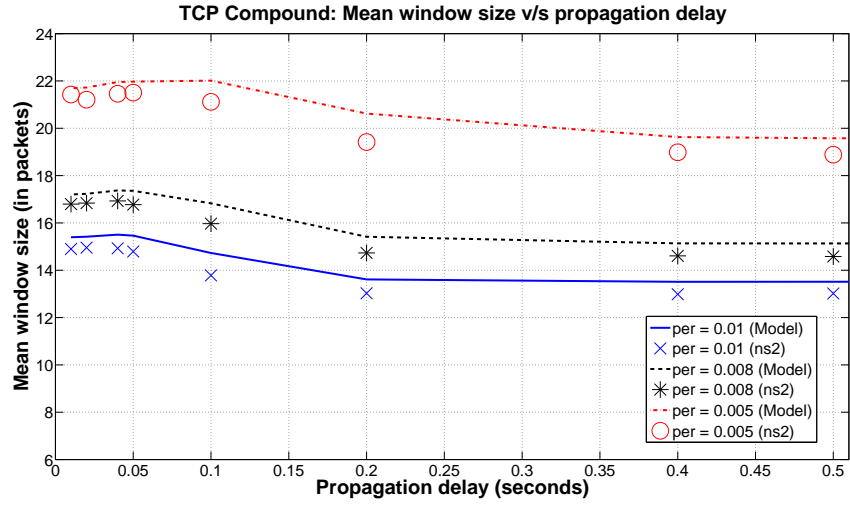


Figure 11: TCP Compound: Effect of change in BDP on $E[W]$.

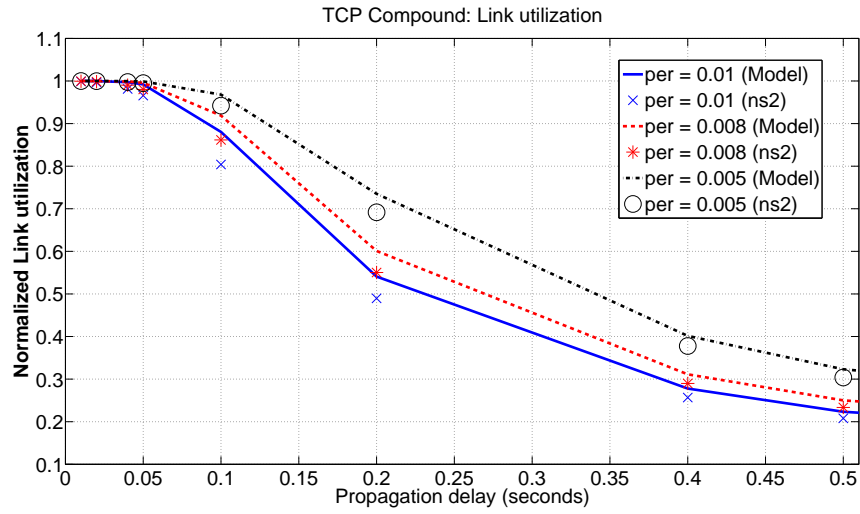


Figure 12: TCP Compound: link utilization.

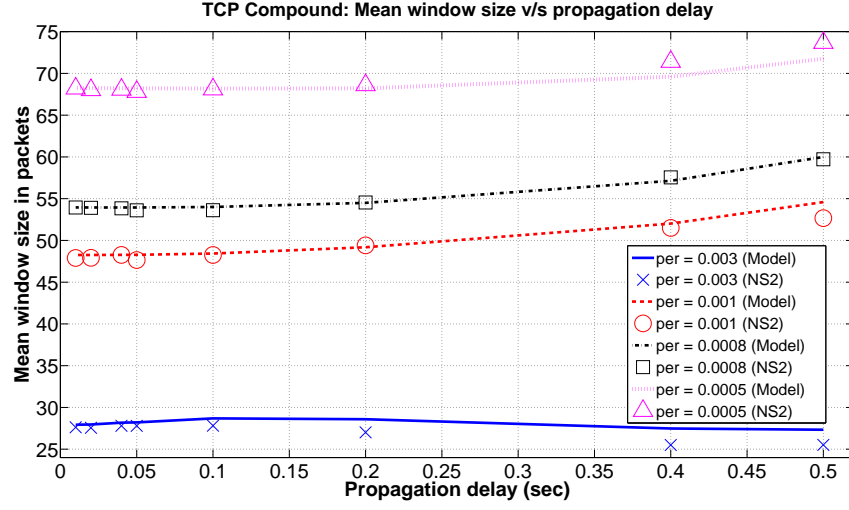


Figure 13: TCP Compound: Effect of change in BDP on $E[W]$.

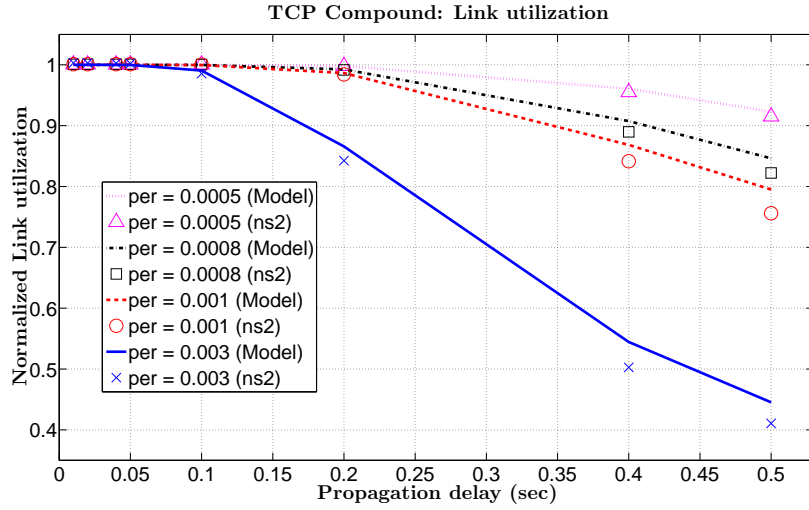


Figure 14: TCP Compound: link utilization.

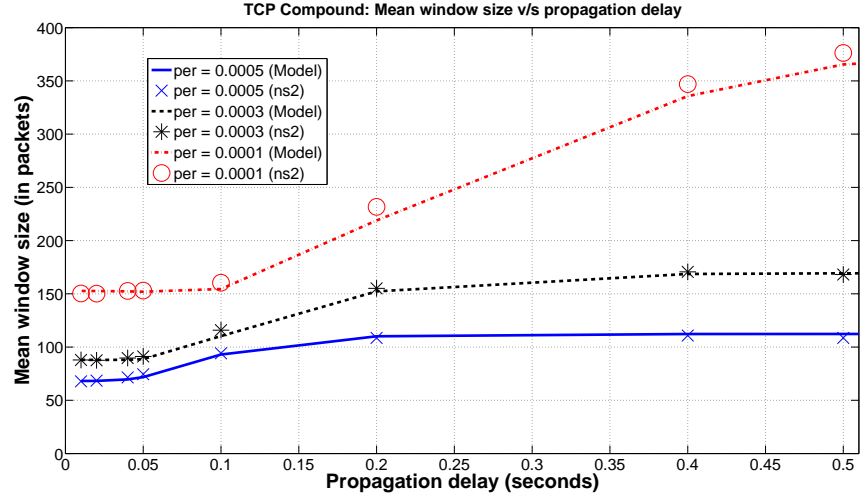


Figure 15: TCP Compound: Effect of change in BDP on $E[W]$.

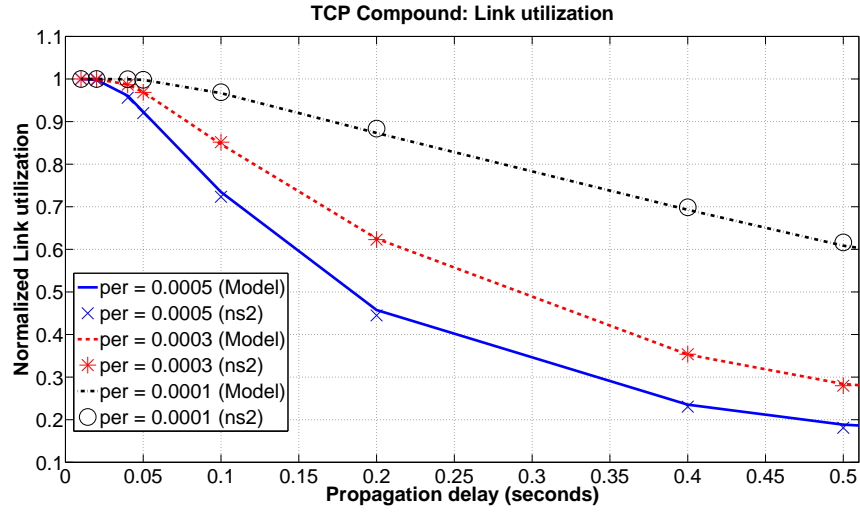


Figure 16: TCP Compound: link utilization.

5 Computational Complexity

We note that Markovian models for TCP CUBIC and TCP Compound are also developed in [35] and [37]. These papers look at the window sizes at the drop epochs and show that this process forms a discrete time Markov chain. Our approach looks at the window sizes at every RTT. The advantage of our approach is that it has reduced complexity. In our model every single-step transition of the Markov chain can lead to one of at most two states since a loss event or no-loss event can give us two potential next states. However if we consider the window size at drop epochs, any state in the state space could potentially be the next state in a single step transition.

Assuming that the TCP Compound Markov chain in [37] has N states, computation of the steady state probabilities, π by use of iteration on $\pi = \pi P$ (where P is the transition probability matrix), requires $O(N^2)$ steps (per iteration) for the model in [37], whereas with our model it reduces to $O(N)$ steps.

For the TCP CUBIC model, assume that $W_{max} \leq N$. Then computation of the steady state probabilities, requires $O(N^2)$ steps (per iteration) for the model in [35]. The Markov model that we have developed for TCP CUBIC requires $O(MN)$ steps (per iteration) where M is the maximum value that T_n in equation (2) takes. When window sizes are large, the TCP CUBIC mode in (2) is dominant. In this case $\max\{T_n\}$ is $O(N^{\frac{1}{3}})$. Thus the Markov model for TCP CUBIC requires $O(N^{\frac{4}{3}})$ steps (per iteration). Hence our Markov models have lesser computational requirements. This can significantly help in computing the stationary distributions and $\mathbb{E}[W]$ for these flows because N can be quite large. In [3] and [6], the average window sizes go up to 10^6 which would require W_{max} to be of that order. In such cases, the computational savings will be huge.

6 Multiple TCPs Through a General Network

In this section we use the models for a single TCP connection for TCP CUBIC and TCP Compound, that we described in Sections 3 and 4 respectively, to study the behaviour of these TCP variants in the general network of Section 2 with multiple bottleneck links. We will also include TCP New-Reno connections for which the theoretical models are available in [21], [20]. Now, the probability of error p_r for flow r will represent the end-to-end probability of packet errors which is seen by the TCP connection r . This may be the cumulative effect of packet overflows in the network and channel errors on wireless links on the route of flow r . We will use two techniques to include the effect of different interacting queues to obtain the throughput of different TCP flows. The first technique approximates the behaviour of connecting links by M/G/1 queues and the second uses an optimization approach.

6.1 The M/G/1 Approximation Method

The main effect different TCP flows passing through the network have on each other is through the queuing delays they cause to each other which then affect their end-to-end RTT. We now describe an approximation to estimate queuing at the bottleneck links.

For a flow $r \in \mathcal{R}$, let $\mathbb{E}[W_r]$ be its average window size, λ_r be its throughput (in packets/sec), $\mathbb{E}[R_r]$ its average RTT (including queuing delays) and $\mathbb{E}[s_r]$ the mean packet length in bits. We denote the average queue length (in packets, of all connections passing through it, excluding the one being serviced) at link $l \in \mathcal{L}$ by $\mathbb{E}[Q_l]$ and the capacity of the link by C_l (in bps). The average RTT for flow r is given by Little's law, as

$$\begin{aligned} \mathbb{E}[R_r] = \Delta_r + & \left(\sum_{l:A(l,r)=1} \frac{\mathbb{E}[Q_l]}{\sum_{r':A(l,r')=1} \lambda_{r'}} \right) \\ & + \left(\sum_{l:B(l,r)=1} \frac{\mathbb{E}[Q_l]}{\sum_{r':A(l,r')=1} \lambda_{r'}} \right). \end{aligned} \quad (12)$$

The first term in RHS of (12) is the constant delay (propagation delay and transmission time), the second term denotes the sum of mean sojourn times for TCP packets on its route and the third term accounts for the total mean sojourn time for the ACK packets on their route. We will assume that the ACKs do not contribute to the queuing, i.e., queuing at the bottleneck links is only caused by TCP packets. This is a reasonable assumption since typically ACK packets are much smaller than TCP packets.

We approximate the behaviour of the links by M/G/1 queues. The mean sojourn time at a link can then be found using the Pollaczek-Khinchine formula [44]. The average queue length, at link l is

$$\mathbb{E}[Q_l] = \frac{(\lambda_l)^2 \mathbb{E}[s_l^2]}{2C_l^2(1 - \rho_l)}, \quad (13)$$

where λ_l is the overall arrival rate at link l , $\mathbb{E}[s_l^2]$ is the second moment of the packet lengths at link l and ρ_l is the utilization factor of link l :

$$\lambda_l = \sum_{r:A(l,r)=1} \lambda_r, \quad (14)$$

$$\mathbb{E}[s_l^2] = \sum_{r:A(l,r)=1} \frac{\lambda_r}{\sum_{r':A(l,r')=1} \lambda_{r'}} \mathbb{E}[s_r^2], \quad (15)$$

$$\rho_l = \sum_{r:A(l,r)=1} \frac{\lambda_r \mathbb{E}[s_r]}{C_l}. \quad (16)$$

The mean window size $\mathbb{E}[W_r]$ for connection r , its throughput, λ_r and its RTT, $\mathbb{E}[R_r]$ are related by Little's law as,

$$\lambda_r = \frac{(1 - p_r) \mathbb{E}[W_r]}{\mathbb{E}[R_r]}. \quad (17)$$

We solve (12), (13) and (17) simultaneously for the unknowns $\bar{\lambda} = \{\lambda_r : r \in \mathcal{R}\}$, $\{\mathbb{E}[W_r] : r \in \mathcal{R}\}$, $\{\mathbb{E}[R_r] : r \in \mathcal{R}\}$ and $\{\mathbb{E}[Q_l] : l \in \mathcal{L}\}$ using (5) for TCP CUBIC, (11) for TCP Compound and

$$E[W_{reno}] = \frac{1.31}{\sqrt{p}} \quad (18)$$

for TCP New Reno from [20].

We illustrate our approach briefly in Algorithm 1. In Algorithm 1, we iteratively find solution to the system of non-linear equations, (12)–(17) using Broyden’s algorithm [45] which is an efficient well known quasi-Newton method. The matrix J is the Jacobian matrix of the fixed point equation, denoted by f . The parameter t is identified using binary search along direction d . The average RTT, $\{\mathbb{E}[R_r] : r \in \mathcal{R}\}$ of the flows are computed using (12). The average window sizes are computed using results from Sections 3 and 4. The procedure gives us the TCP performance measures, $\{\mathbb{E}[W_r], \lambda_r, \mathbb{E}[R_r] : r \in \mathcal{R}\}$.

The M/G/1 approximation technique we describe here is an approximation as the arrival process for TCP packets at a queue is not Poisson. This approximation helps us to easily compute the queuing delay at each link without prior identification of the bottleneck links. However the TCP dynamics have also been captured in equation (17) via mean window lengths. Through simulation results described in Section 7, we will see that these simplifying approximations are reasonably accurate.

Algorithm 1 Throughput via M/G/1 approximation.

INPUT: $\{p_r, \Delta_r, \mathbb{E}[s_r] : r \in \mathcal{R}\}, \{C_l : l \in \mathcal{L}\}, A, B, \epsilon$

OUTPUT: $\{\lambda_r, \mathbb{E}[W_r], \mathbb{E}[R_r] : r \in \mathcal{R}\}$

Initialize $\bar{\lambda} = \{\lambda_r : r \in \mathcal{R}\}$

Compute $\{\mathbb{E}[R_r](\bar{\lambda}), \mathbb{E}[W_r](\bar{\lambda}) : r \in \mathcal{R}\}$

Compute

$$f(\bar{\lambda}) = \left\{ \lambda_r - \frac{(1 - p_r)\mathbb{E}[W_r](\bar{\lambda})}{\mathbb{E}[R_r](\bar{\lambda})} \right\}_{r \in \mathcal{R}}$$

Compute Jacobian matrix, J for f

$H = J^{-1}$

while $\|f(\bar{\lambda})\|_2 > \epsilon$ **do**

 Set $d = -Hf(\bar{\lambda})$

 Choose t such that $\|f(\bar{\lambda} + dt)\|$ is minimized

 Set $\bar{\lambda} = \bar{\lambda} + dt$

 Compute $\{\mathbb{E}[R_r](\bar{\lambda}), \mathbb{E}[W_r](\bar{\lambda}) : r \in \mathcal{R}\}$

 Recompute H using [45]

end while

return $\{\mathbb{E}[W_r], \lambda_r, \mathbb{E}[R_r] : r \in \mathcal{R}\}$

6.2 The Optimization Approach

In this section, we describe an optimization approach to compute the average throughput in the network. We will use the expressions (5) for TCP CUBIC, (11) for TCP Compound and (18) for TCP Reno and use them in an optimization program. This approach is similar to the one used in [30]. However, [30] assumes that there is negligible queuing in the network and has only TCP Reno whereas in our model the queuing may be non-negligible and also has TCP Compound and TCP CUBIC connections.

We will first consider a simple approximation when there is only one bottleneck queue. Suppose there are multiple TCP connections going through a single bottleneck link router. The TCP throughput for the different connections can be computed using the following approximation. Let C be the bottleneck link capacity (in bps) and let D be the average queuing delay at the bottleneck link. We can compute the average window size for TCP CUBIC, TCP Compound and New Reno using equations (5), (11) and (18) respectively. If we have $\sum_r \frac{(1-p_r)\mathbb{E}[W_r]\mathbb{E}[s_r]}{\Delta_r} \leq C$, we set the throughput (in packets/sec) for connection r as $\lambda_r = \frac{(1-p_r)\mathbb{E}[W_r]}{\Delta_r}$ and $M = 0$, else we find M such that

$$\sum_r \frac{(1-p_r)\mathbb{E}[W_r]\mathbb{E}[s_r]}{\Delta_r + M} = C. \quad (19)$$

In the above case we have assumed that the bottleneck link is either operating at full capacity or has zero queuing. Extending this for the multihop network, we assume that if $\sum_{r:A(l,r)=1} \lambda_r \mathbb{E}[s_r] < C_l$ then $\mathbb{E}[Q_l] = 0$. Therefore, for each $r \in \mathcal{R}$ and each $l \in \mathcal{L}$, we have

$$(1-p_r)\mathbb{E}[W_r] = \lambda_r(\Delta_r + \sum_{l:A(l,r)=1} M_l), \quad (20)$$

and

$$M_l(C_l - \sum_{r:A(l,r)=1} \lambda_r \mathbb{E}[s_r]) = 0, \quad (21)$$

where M_l is the mean sojourn time at link l . The equation (20) comes from the Little's law, whereas (21) is a restatement of our assumption. Besides (20) and (21), we require

$$\sum_{r:A(l,r)=1} \lambda_r \mathbb{E}[s_r] \leq C. \quad (22)$$

Assuming $\mathbb{E}[W_r]$ is known, we can solve (20)–(22) (for the unknowns $\{\lambda_r \in \mathcal{R}\}$ and $\{M_l \in \mathcal{L}\}$) using Algorithm 2. The algorithm stops if the successive approximations to M^* are ϵ -close. The parameter d in the algorithm is the step size which dictates the speed of convergence.

The average window size expressions (as computed using the Markov chain models in Sections 3 and 4) are a function of the average RTT (for TCP CUBIC) and average queue size (for TCP Compound) of the connections. Thus,

Algorithm 2 Solution to (20)–(22).

INPUT: $\{\mathbb{E}[W_r], p_r, \Delta_r, \mathbb{E}[s_r] : r \in \mathcal{R}\}, \{C_l : l \in \mathcal{L}\}, A, \epsilon$.

OUTPUT: $M^* = \{M_l^* : l \in \mathcal{L}\}, \lambda^* = \{\lambda_r^* : r \in \mathcal{R}\}$.

Initialize $M^1 = \{M_l^1 : l \in \mathcal{L}\}$.

Initialize $M^2 = \{M_l^2 : l \in \mathcal{L}\}$.

Set step size d .

repeat

$M^1 = M^2$.

for all $r \in \mathcal{R}$ **do**

$$\lambda_r = \frac{(1-p_r)\mathbb{E}[W_r]}{\Delta_r + \sum_{j:A(j,r)=1} M_j}.$$

end for

for all $l \in \mathcal{L}$ **do**

$$M_l^2 = (M_l^1 - d(C_l - \sum_{r:A(l,r)=1} \lambda_r \mathbb{E}[s_r]))^+$$

end for

until $\|M^1 - M^2\| \leq \epsilon$

$M^* = M^2$

$\lambda^* = \{\lambda_r : r \in \mathcal{R}\}$

return M^*, λ^* .

we need to solve for the average window size equations, and equations (20) and (21) simultaneously. As in Section 6.1, we use Broyden's algorithm for simultaneously solving these equations and obtaining the throughput of the different connections. We describe our approach in brief in Algorithm 3. In the M/G/1 approach, we initialize Algorithm 1 with a guess of the throughput of the different flows. Here, we initialize Algorithm 3, with a guess of \bar{X} which consists of $\{\mathbb{E}[R_r] : r \in \mathcal{R} \text{ and } r \text{ is a CUBIC connections}\}$ and $\{\mathbb{E}[Q_r] : r \in \mathcal{R} \text{ and } r \text{ is a Compound connections}\}$ so that we can compute $\{\mathbb{E}[W_r] : r \in \mathcal{R}\}$. We then compute the throughputs, $\{\lambda_r : r \in \mathcal{R}\}$, of the different connections and the average queue sizes, $\{M_l : l \in \mathcal{L}\}$, using Algorithm 2. Since in this approach, we did not account for the queuing incurred by the ACK packets, we need to update, Δ_r and hence $\{\mathbb{E}[R_r] : r \in \mathcal{R}\}$ iteratively using the queuing estimates $\{M_l : l \in \mathcal{L}\}$. We then iteratively find solution to the equation $g(\bar{X}) = 0$ (see (23)) using Broyden's algorithm [45]. The procedure gives us the TCP performance measures, $\{\mathbb{E}[W_r], \lambda_r, \mathbb{E}[R_r] : r \in \mathcal{R}\}$.

The optimization problem, provided in Proposition 1 is another way to solve (20)–(22). In [29], a similar optimization approach was used to obtain the rates attained by congestion control algorithms with fixed end-to-end window sizes in a network of queues which use FIFO scheduling. We use the following optimization program for different TCP variants with varying window sizes with finite means.

Proposition 1. *In the general system model, under the above assumption, (i.e.,*

Algorithm 3 Throughput via Optimization Approach.

INPUT: $\{p_r, \Delta_r, \mathbb{E}[s_r] : r \in \mathcal{R}\}, \{C_l : l \in \mathcal{L}\}, A, B, \epsilon$
 OUTPUT: $\{\lambda_r, \mathbb{E}[W_r], \mathbb{E}[R_r] : r \in \mathcal{R}\}$
 Initialize $\overline{X} = \{\mathbb{E}[R_s], \mathbb{E}[Q_t] : s, t \in \mathcal{R}\}$, where connections s, t index TCP
 CUBIC and TCP Compound connections respectively.
 Compute $\{\mathbb{E}[W_r](\overline{X}) : r \in \mathcal{R}\}$
 Compute $\{\lambda_r : r \in \mathcal{R}\}, \{M_l : l \in \mathcal{L}\}$ using Algorithm 2
 Update $\{\Delta_r, \mathbb{E}[R_r] : r \in \mathcal{R}\}$ using $\{M_l : l \in \mathcal{L}\}$
 Compute

$$g(\overline{X}) = \left\{ \lambda_r(\overline{X}) - \frac{(1 - p_r)\mathbb{E}[W_r](\overline{X})}{\mathbb{E}[R_r]} \right\}_{r \in \mathcal{R}} \quad (23)$$

Compute Jacobian matrix, J for g
 $H = J^{-1}$
while $\|g(\overline{X})\|_2 > \epsilon$ **do**
 Set $d = -Hf(\overline{X})$
 Choose t such that $\|g(\overline{X} + dt)\|$ is minimized
 Set $\overline{X} = \overline{X} + dt$
 Compute $\{\mathbb{E}[W_r](\overline{X}) : r \in \mathcal{R}\}$
 Compute $\{\lambda_r : r \in \mathcal{R}\}$ using Algorithm 2
 Update $\{\Delta_r, \mathbb{E}[R_r] : r \in \mathcal{R}\}$ using $\{M_l : l \in \mathcal{L}\}$
 Recompute H using [45]
end while
return $\{\mathbb{E}[W_r], \lambda_r, \mathbb{E}[R_r] : r \in \mathcal{R}\}$

if $\sum_{r:A(l,r)=1} \lambda_r < C_l$ then $\mathbb{E}[Q_l] = 0$) the primal solutions of the optimization problem

$$\max \sum_{r \in \mathcal{R}} \mathbb{E}[s_r] \left((1 - p_r) \mathbb{E}[W_r] \log(\lambda_r) - \lambda_r \Delta_r \right) \quad (24)$$

such that

$$\lambda_r \geq 0, \forall r \in \mathcal{R} \text{ and } \sum_{r:A(l,r)=1} \lambda_r \mathbb{E}[s_r] \leq C_l, \forall l \in \mathcal{L},$$

provide the system throughputs, $\{\lambda_r : r \in \mathcal{R}\}$.

Proof. Consider the optimization problem (24). Let $\{\lambda_r^* : r \in \mathcal{R}\}$ denote the optimal point for (24) and let us denote the dual optimal points corresponding to the capacity constraints by $\{M_l^* : l \in \mathcal{L}\}$ and the dual optimal points corresponding to the non-negativity constraints by $\{\gamma_r^* : r \in \mathcal{R}\}$. The KKT conditions [46] for the above problem are given as follows. For each $r \in \mathcal{R}$ and each $l \in \mathcal{L}$, we have

$$(1 - p_r) \mathbb{E}[W_r] = \lambda_r^* \left(\Delta_r + \sum_{l:A(l,r)=1} M_l^* + \frac{\gamma_r^*}{\mathbb{E}[s_r]} \right), \quad (25a)$$

$$M_l^* (C_l - \sum_{r:A(l,r)=1} \lambda_r^* \mathbb{E}[s_r]) = 0, \quad (25b)$$

$$\gamma_r^* \lambda_r^* = 0. \quad (25c)$$

Also $M_l^* \geq 0$, for each l and $\lambda_r^* \geq 0$, for each r . The objective function is concave and $\lambda_r = 0, \forall r \in \mathcal{R}$ is strictly feasible. Therefore by Slater's condition [46], strong duality holds. Hence any pair of primal optimal and dual optimal points must satisfy the KKT conditions listed in (25). Since the problem is concave, the KKT conditions are also sufficient for optimality. Therefore we conclude that any solution to (24) also satisfies (25) and vice versa.

For each $r \in \mathcal{R}$, the term $\mathbb{E}[W_r]$ in (20) is greater than 0 (since TCP window sizes are ≥ 1). Thus, for any solution to (20)–(22), $\lambda_r > 0$ for each $r \in \mathcal{R}$. Therefore under the above assumptions, any $\{\lambda_r > 0 : r \in \mathcal{R}\}$ and $\{M_l \geq 0 : l \in \mathcal{L}\}$ which satisfies (20) and (21) satisfies the KKT conditions (25) with $\gamma_r^* = 0$ for all $r \in \mathcal{R}$. Therefore λ_r for all $r \in \mathcal{R}$ and M_l for all $l \in \mathcal{L}$ which satisfy (20), (21) and (22) also satisfy the KKT conditions. \square

The optimization program (24) is a strict concave optimization problem with affine constraints and hence has a unique solution. Thus this optimization problem can be solved via the usual convex-optimization algorithms [46]. We note that as strong duality holds for (24), we can solve it via its dual. Algorithm 2 solves the optimization program (24) via $\{M_l : l \in \mathcal{L}\}$, the dual variables. It is a projected gradient method used to minimize the dual of optimization program (24). In Algorithm 2, the dual variable, M_l , i.e., the queuing delay at link l is reduced whenever the arrival rate $\sum_{r:A(l,r)=1} \lambda_r \mathbb{E}[s_r]$ at link l exceeds capacity

C_l and is increased whenever the arrival rate $\sum_{r:A(l,r)=1} \lambda_r \mathbb{E}[s_r]$ at link l is less than capacity C_l . If we choose d small enough, Algorithm 2 converges after either setting M_l close to 0 or when $\sum_{r:A(l,r)=1} \lambda_r \mathbb{E}[s_r]$ at link l almost equals the capacity C_l , for all $l \in \mathcal{L}$. Similar techniques for solving optimization programs have been used in [47] and [48]. Through simulation results described in Section 7, we will see that our approximations made in this approach are reasonable.

7 Simulation Results

In this section, we will validate our models using ns2 simulations. We use extensions from [49] for high speed TCP which use actual Linux TCP code for the simulations. We first look at the results for single bottleneck link.

7.1 Single Bottleneck Link

We first consider the case when multiple flows go through a single bottleneck link and the remaining connecting links have sufficiently high capacity. We use the M/G/1 approximation based model and the optimization approach for a single bottleneck queue discussed earlier in Sections 6.1 and 6.2 respectively. In this setup, we have 12 flows (4 Compound, 4 CUBIC, 4 New Reno) going through a single bottleneck link. We number the flows 1–12, the flows indexed i such that $\text{mod}(i, 3) = 1$ ¹ are TCP Compound flows, $\text{mod}(i, 3) = 2$ are TCP CUBIC and the rest are TCP New Reno. These flows have different propagation delays. The first three flows have propagation delay of 0.01 sec, flows 4–6 have delay of 0.02 sec, flows 7–9 have delay of 0.1 sec and the last three flows have propagation delay of 0.2 sec. All flows have TCP packet size of 1050 bytes which is the default packet size of ns2. The packet error rates for all flows is 0.001. We consider three cases with different bottleneck link capacities. In Figure 17, the bottleneck capacity is 50 Mbps which illustrates the situation when there is severe congestion. Figures 18, 19 have bottleneck capacities of 100 Mbps and 1 Gbps illustrating moderate and low congestion scenario respectively. Our theoretical results are quite close to the simulation results and all errors are less than 10%. In Table 1, we compare results for the average queuing at the bottleneck link and the bottleneck link utilization.

From the Figures 17, 18 and 19, we see that among the different TCP versions with same propagation delay, TCP Compound (flows with $\text{mod}(i, 3) = 1$) gets the highest throughput when the propagation delay is small ($\Delta \leq 0.1$ sec) whereas TCP CUBIC (flows with $\text{mod}(i, 3) = 2$) gets the highest throughput when propagation delay is large ($\Delta = 0.2$ sec). However as link speeds reduce from 1 Gbps to 50 Mbps we see that the different TCP versions have a more equitable share of the bottleneck link capacity, for $\Delta \leq 0.1$ sec. This happens due to the increase in RTT, due to queuing, when the link speed is reduced from

¹ $\text{mod}(i, j) = i \% j$

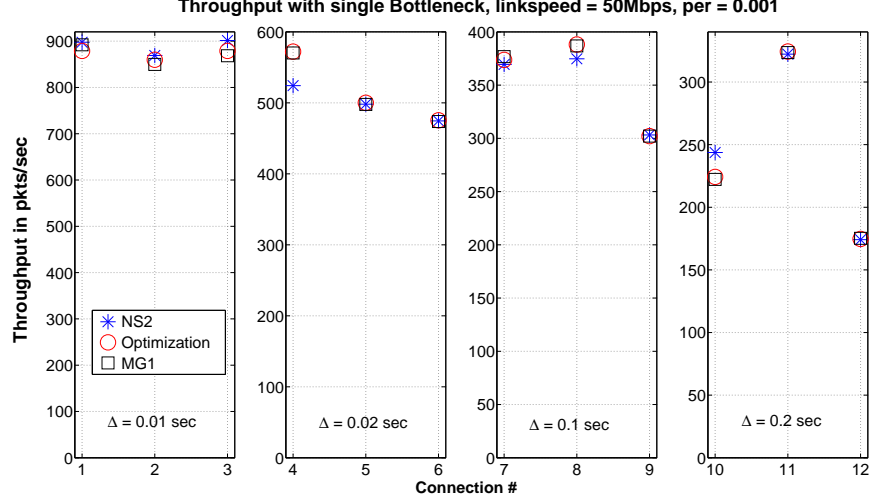


Figure 17: Single bottleneck with link capacity 50 Mbps.

Table 1: Single bottleneck: average queue size, link utilization at bottleneck queue.

Link Speed	Average Queue Size (in packets)			Normalized Link utilization		
	ns2	Approx. Model	MG1	ns2	Approx. Model	MG1
50Mbps	223.2	220.2	217.4	1.0	1.0	0.997
100Mbps	92.4	94.0	87.3	0.999	1.0	0.994
1Gbps	0.17	0.0	0.016	0.159	0.166	0.166

1 Gbps to 50 Mbps which causes (a) increase in average window size of TCP CUBIC as for fixed PER its window size grows with RTT and (b) decrease in average window size of TCP Compound which decreases with increase in queue size.

7.2 Multiple Bottleneck Links

In this section, we consider examples with multiple bottleneck links. We consider the case when besides TCP packets, ACKs also get delayed. In the first example in this scenario, we have ten routers and a total of 15 flows, 5 using TCP Compound, 5 using CUBIC and 5 using Reno. We group the flows into 5 flow groups, $(F_1 - F_5)$ with each group consisting of 3 flows with one flow each of TCP Compound, CUBIC and Reno. The network topology is shown in Figure 20. The links are bidirectional and are symmetric, i.e., forward direction and reverse direction have the same bandwidth and propagation delay. We note that in this example, the ACKs may also be subject to queuing which is accounted for in both the M/G/1 based models and the optimization based discussed in Sections 6.1 and 6.2.

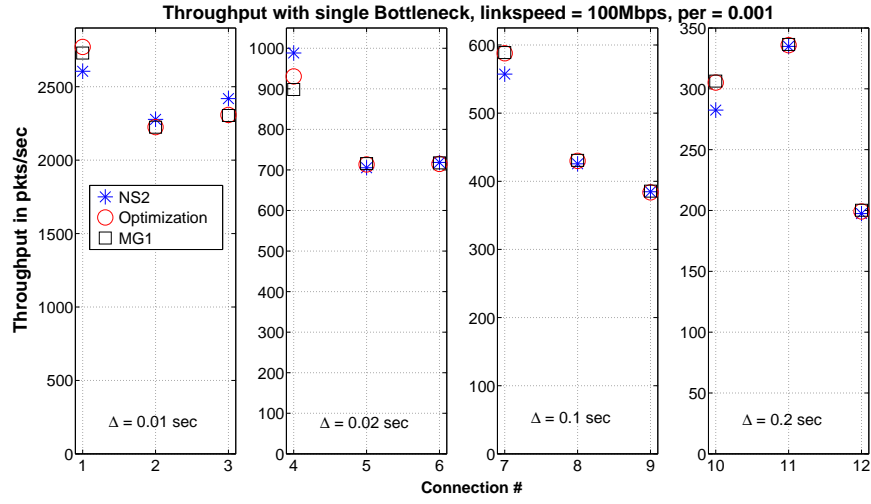


Figure 18: Single bottleneck with link capacity 100 Mbps.

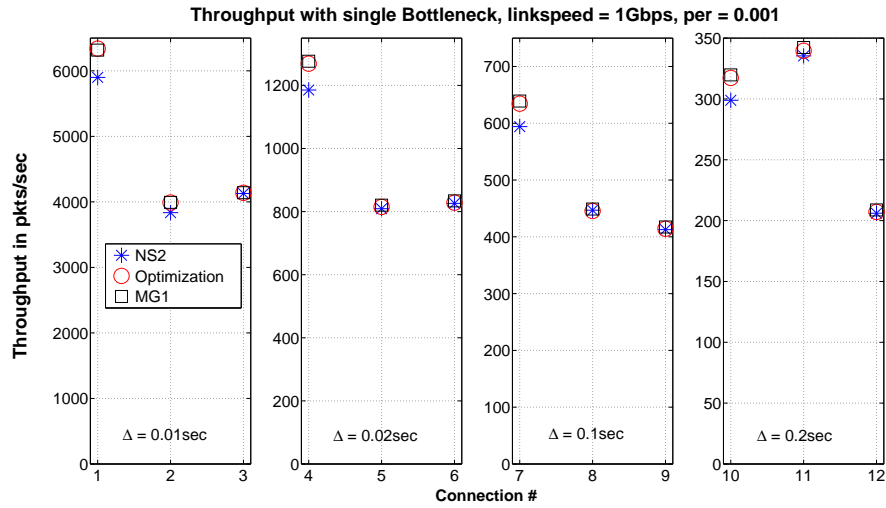


Figure 19: Single bottleneck with link capacity 1 Gbps.

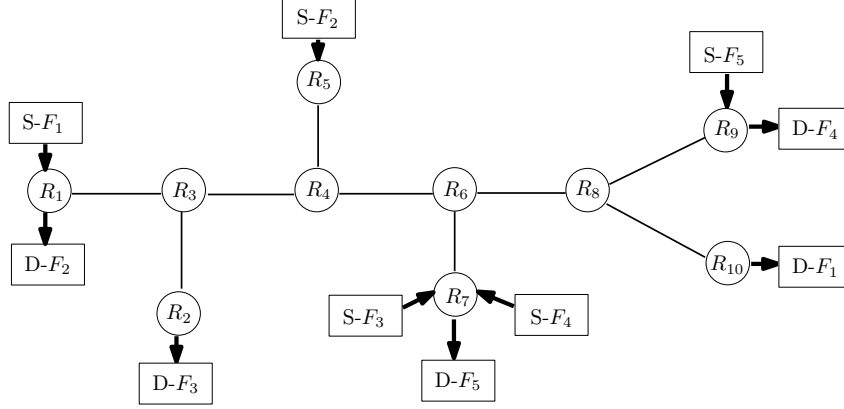


Figure 20: Topology of scenario with 10 Routers, 15 Flows.

In Figures 21 and 23 we compare the throughputs obtained by the different flows for two different configurations with network topology given in Figure 20. In Figures 22 and 24 we compare the link utilization for links with link utilization $> 70\%$ (in ns2 simulations) in these two configurations. The two configurations have different link speeds (mentioned in Figures 22 and 24) and we denote them as config. 1 and config. 2. All flows in a flow group have the same propagation delay and packet error rate. Flows are numbered 1 – 15. Flows numbered i such that $\text{mod}(i, 3) = 1$ are TCP Compound flows, flows with $\text{mod}(i, 3) = 2$ are TCP CUBIC whereas rest are TCP New Reno flows. The parameters used for simulation are mentioned in Figures 21 and 23. The link speeds are mentioned in Mbps and the propagation delays are mentioned in sec. All flows have packet sizes 1050 bytes which is the default value in ns2. For the first configuration, for the throughput obtained by the different flows, the M/G/1 model results differ from the ns2 simulations by less than 8% whereas for the optimization approach the maximum difference is 10.5%. In the second configuration, for the throughput obtained by the different flows, both models differ from the ns2 simulations by less than 8%. From 22 and 24, we see that both the M/G/1 and the optimization techniques identify the bottleneck links (links with utilization $> 70\%$) correctly. The difference in link utilization, as compared to ns2 simulations, for all the links in the network, is less than 4% for the M/G/1 technique in both the configurations. For the optimization technique, the difference from ns2 simulations is less than 7%.

Next we consider the scenario with 12 routers and 18 flows with the topology shown in Figure 25. We now have 6 groups of flows with each group consisting of 3 flows with one flow each of TCP Compound, CUBIC and Reno. In Figures 26 and 28 we compare the throughputs obtained by the different flows for two different configurations with network topology given in Figure 25. In Figures 27 and 29 we compare the link utilization for links with link utilization $> 70\%$ (in ns2 simulations). The configurations differ in the link speeds (mentioned

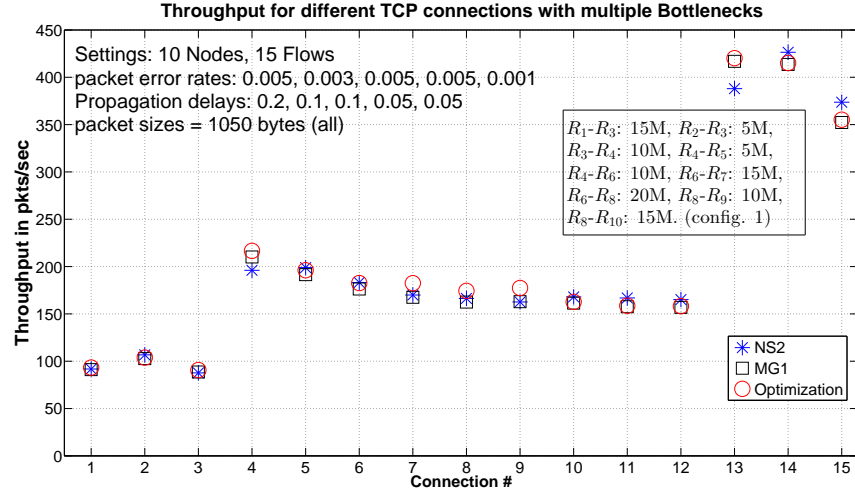


Figure 21: Throughputs for example in Figure 20, config. 1.

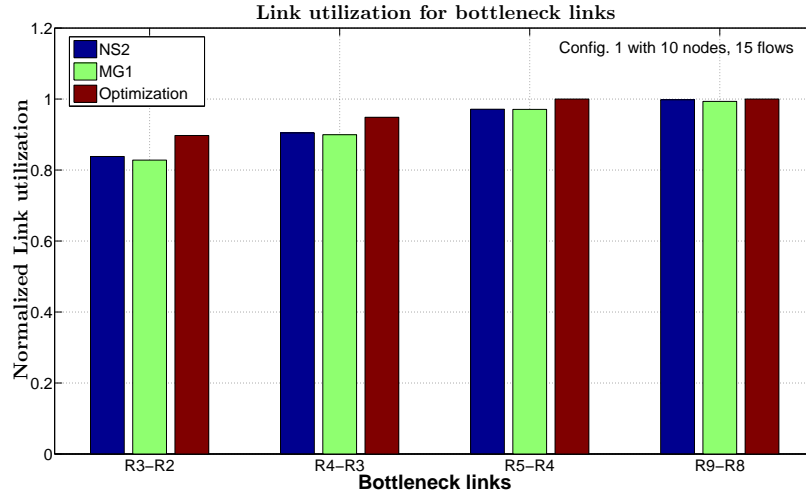


Figure 22: Link utilization for example in Figure 20, config. 1.

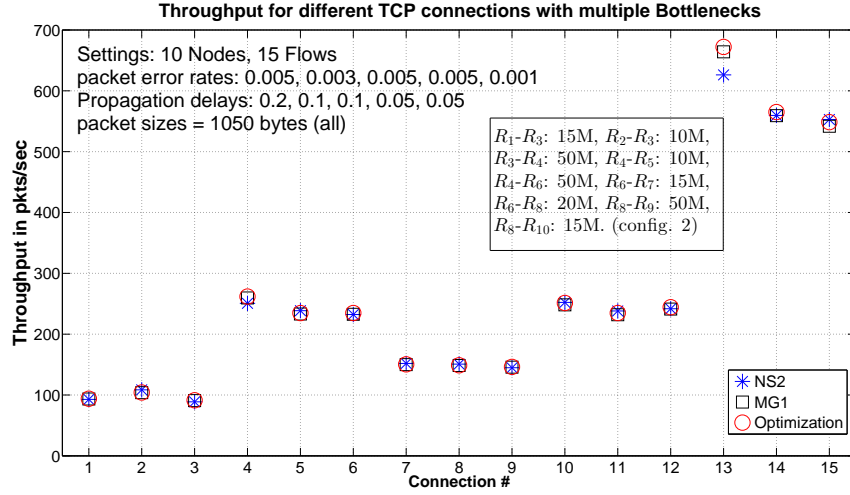


Figure 23: Throughputs for example in Figure 20, config. 2.

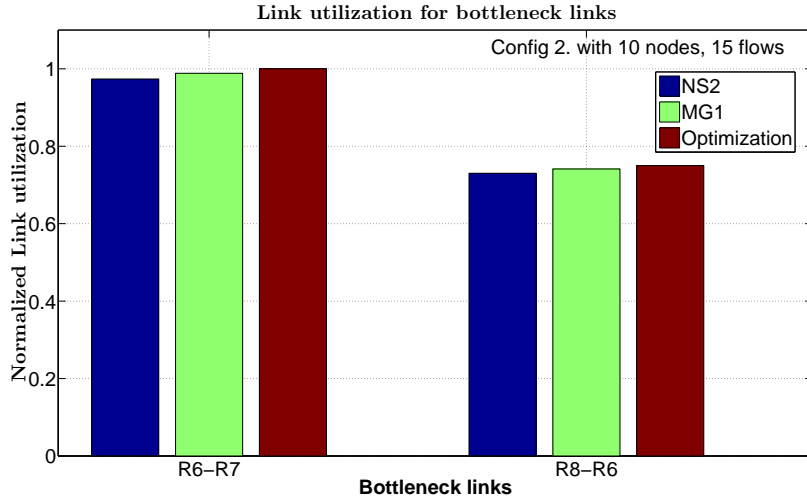


Figure 24: Link Utilization for example in Figure 20, config. 2.

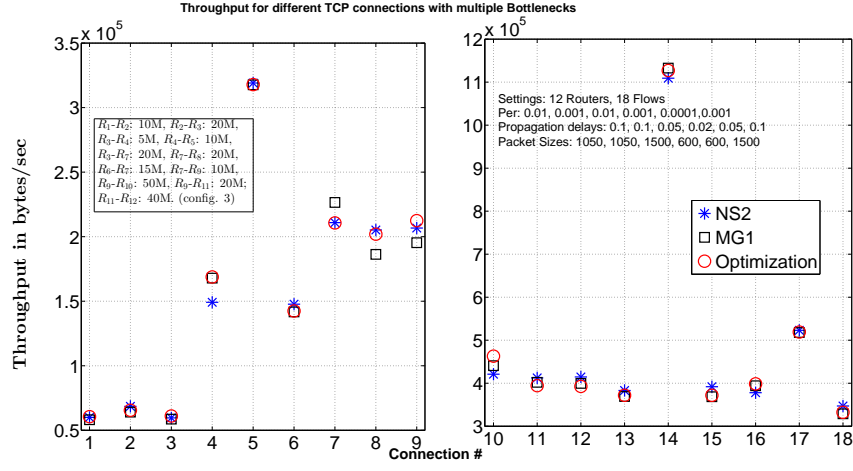


Figure 26: Throughputs for example in Figure 25, config. 3.

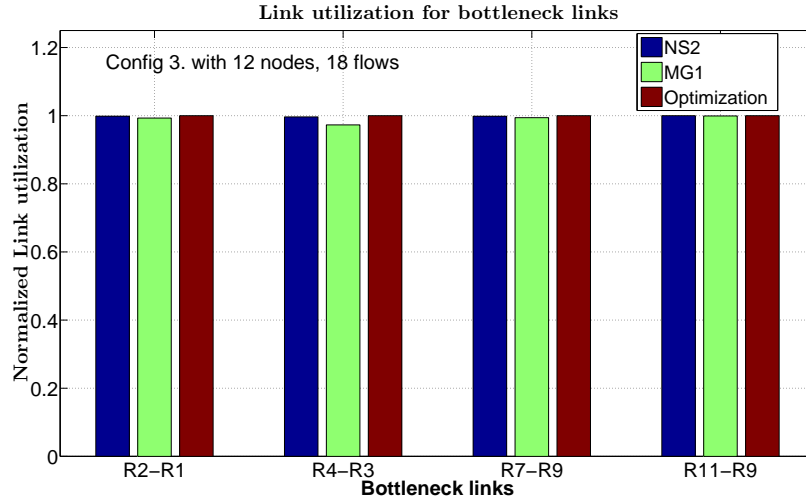


Figure 27: Link Utilization for example in Figure 25, config. 3.

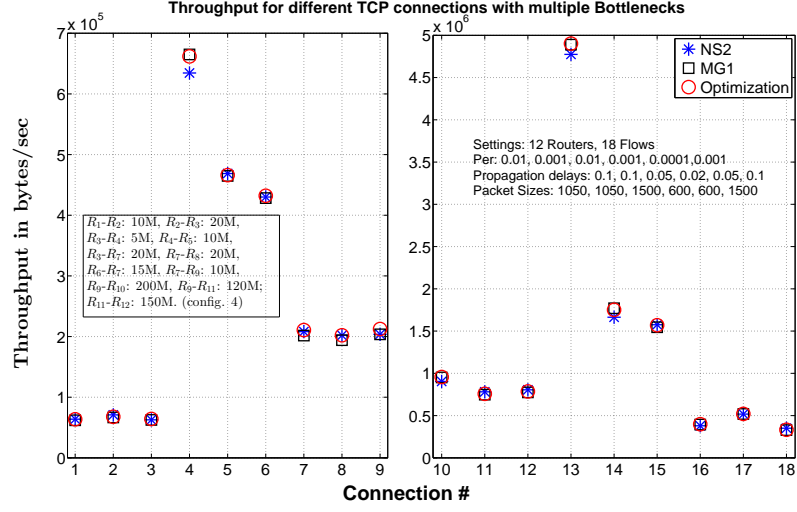


Figure 28: Throughputs for example in Figure 25, config. 4.

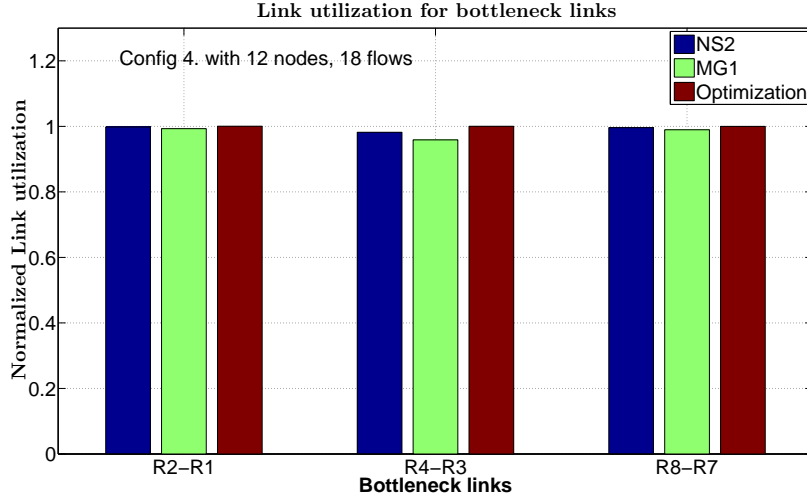


Figure 29: Link Utilization for example in Figure 25, config. 4.

pound, we compute the average window size for a single TCP flow with non-negligible queuing and random losses. We use two techniques to compute the steady state throughput for multiple TCP flows (which could be using TCP CUBIC, TCP Compound, TCP New Reno) going through a multihop network. The first technique approximates the links by M/G/1 queues. The second technique uses an optimization program whose solution approximates the steady state throughput for the different TCP flows. We compare results obtained from our techniques with ns2 simulation. Our results match well with simulations.

References

- [1] G. Huston, “Gigabit TCP,” *Internet Protocol Journal*, 2006.
- [2] D.-M. Chiu and R. Jain, “Analysis of the increase and decrease algorithms for congestion avoidance in computer networks,” *Computer Networks and ISDN systems*, vol. 17, no. 1, pp. 1–14, 1989.
- [3] S. Ha, I. Rhee, and L. Xu, “CUBIC: a new TCP-friendly high-speed TCP variant,” *SIGOPS Oper. Syst. Rev.*, vol. 42, pp. 64–74, July 2008.
- [4] C. Jin, D. Wei, S. H. Low, J. Bunn, H. D. Choe, J. C. Doyle, H. Newman, S. Ravot, and S. Singh, “Fast TCP: From theory to experiments,” *IEEE Network*, vol. 19, pp. 4–11, 2005.
- [5] D. Leith and R. Shorten, “H-TCP: TCP for high-speed and long-distance networks,” in *Proceedings of PFLDnet*, vol. 2004, 2004.
- [6] K. Tan, J. Song, Q. Zhang, and M. Sridharan, “A Compound TCP Approach for High-Speed and Long Distance Networks,” in *IEEE Infocom*, 2006.
- [7] P. Yang, J. Shao, W. Luo, L. Xu, J. Deogun, and Y. Lu, “TCP Congestion Avoidance Algorithm Identification,” *Networking, IEEE/ACM Transactions on*, vol. 22, no. 4, pp. 1311–1324, Aug 2014.
- [8] S. Floyd, “HighSpeed TCP for Large Congestion Windows,” RFC 3649 (Experimental), Internet Engineering Task Force, December 2003.
- [9] T. Kelly, “Scalable TCP: Improving performance in highspeed wide area networks,” *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 2, pp. 83–91, 2003.
- [10] L. Xu, K. Harfoush, and I. Rhee, “Binary Increase Congestion Control (BIC) for Fast Long-Distance Networks,” in *Infocom ’04*, 2004.
- [11] D. X. Wei, C. Jin, S. H. Low, and S. Hegde, “FAST TCP: motivation, architecture, algorithms, performance,” *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 1246–1259, Dec. 2006.

- [12] L. S. Brakmo and L. L. Peterson, "TCP Vegas: End to end congestion avoidance on a global Internet," *Selected Areas in Communications, IEEE Journal on*, vol. 13, no. 8, pp. 1465–1480, 1995.
- [13] T. Bonald, "Comparison of TCP Reno and TCP Vegas: efficiency and fairness," *Performance Evaluation*, vol. 36-37, pp. 307 – 332, 1999.
- [14] A. Afanasyev, N. Tilley, P. Reiher, and L. Kleinrock, "Host-to-Host Congestion Control for TCP," *Communications Surveys Tutorials, IEEE*, vol. 12, no. 3, pp. 304 –342, quarter 2010.
- [15] S. Jain and G. Raina, "An experimental evaluation of CUBIC TCP in a small buffer regime," in *Communications (NCC), 2011 National Conference on*. IEEE, 2011, pp. 1–5.
- [16] M. Bateman, S. Bhatti, G. Bigwood, D. Rehunathan, C. Allison, T. Henderson, and D. Miras, "A comparison of TCP behaviour at high speeds using ns-2 and linux," in *Proceedings of the 11th communications and networking simulation symposium*, ser. CNS '08. New York, NY, USA: ACM, 2008, pp. 30–37.
- [17] M. C. Weigle, P. Sharma, and J. Freeman IV, "Performance of competing high-speed TCP flows," in *Proceedings of NETWORKING*, Coimbra, Portugal, may 2006, pp. 476–487.
- [18] S. Molnár, B. Sonkoly, and T. A. Trinh, "A comprehensive TCP fairness analysis in high speed networks," *Computer Communications*, vol. 32, no. 13, pp. 1460–1484, 2009.
- [19] L. Xue, S. Kumar, C. Cui, and S.-J. Park, "A study of fairness among heterogeneous TCP variants over 10Gbps high-speed optical networks," *Optical Switching and Networking*, vol. 13, no. 0, pp. 124 – 134, 2014.
- [20] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, "The macroscopic behavior of the TCP congestion avoidance algorithm," *SIGCOMM Comput. Commun. Rev.*, vol. 27, pp. 67–82, July 1997.
- [21] J. Padhye, V. Firoiu, D. F. Towsley, and J. F. Kurose, "Modeling TCP Reno Performance: A Simple Model and Its Empirical Validation," *IEEE//ACM Transactions on Networking*, vol. 8, pp. 133–145, 2000.
- [22] N. Cardwell, S. Savage, and T. Anderson, "Modeling TCP latency," in *IEEE Infocom*, 2000, pp. 1724–1751.
- [23] V. Sharma and A. Gupta, "Performance analysis of routers with TCP and UDP connections with priority and RED control," in *Proceedings of the 15th international conference on Computer communication*, ser. ICC '02. Washington, DC, USA: International Council for Computer Communication, 2002, pp. 1015–1022.

- [24] V. Sharma and P. Purkayastha, “Stability and Analysis of TCP Connections with RED Control and Exogenous Traffic,” *Queueing Syst. Theory Appl.*, vol. 48, pp. 193–235, November 2004.
- [25] A. Gupta and V. Sharma, “A unified approach for analyzing persistent, non-persistent and ON–OFF TCP sessions in the Internet,” *Performance Evaluation*, vol. 63, no. 2, pp. 79–98, 2006.
- [26] S. Kunniyur and R. Srikant, “End-to-end Congestion Control Schemes: Utility Functions, Random Losses and ECN Marks,” *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 689–702, Oct. 2003.
- [27] P. Hurley, J. Yves Le Boudec, and P. Thiran, “A note on the fairness of additive increase and multiplicative decrease,” in *In Proceedings of ITC-16*, 1999, pp. 467–478.
- [28] S. Low, “A duality model of TCP and queue management algorithms,” *Networking, IEEE/ACM Transactions on*, vol. 11, no. 4, pp. 525–536, 2003.
- [29] L. Massoulie and J. Roberts, “Bandwidth sharing: objectives and algorithms,” *Networking, IEEE/ACM Transactions on*, vol. 10, no. 3, pp. 320–328, June 2002.
- [30] E. Altman, K. Avrachenkov, and C. Barakat, “TCP Network Calculus: The case of large delay-bandwidth product,” in *IEEE Infocom*, 2002.
- [31] T. V. Lakshman and U. Madhow, “The performance of TCP/IP for networks with high bandwidth-delay products and random loss,” *IEEE/ACM Trans. Netw.*, vol. 5, pp. 336–350, June 1997.
- [32] V. Misra, W.-B. Gong, and D. Towsley, “Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to RED,” in *ACM SIGCOMM Computer Communication Review*, vol. 30, no. 4. ACM, 2000, pp. 151–160.
- [33] F. Baccelli, D. McDonald, and J. Reynier, “A mean-field model for multiple TCP connections through a buffer implementing RED,” *Performance Evaluation*, vol. 49, no. 14, pp. 77 – 97, 2002.
- [34] S. Shakkottai and R. Srikant, “How good are deterministic fluid models of internet congestion control?” in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, 2002, pp. 497–505 vol.2.
- [35] W. Bao, V. W. S. Wong, and V. C. M. Leung, “A Model for Steady State Throughput of TCP CUBIC,” in *GLOBECOM*, 2010.
- [36] S. Belhareth, L. Sassatelli, D. Collange, D. Lopez-Pacheco, and G. Urvoy-Keller, “Understanding TCP cubic performance in the cloud: A mean-field approach,” in *Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on*, Nov 2013, pp. 190–194.

- [37] A. Blanc, K. Avrachenkov, D. Collange, and G. Neglia, "Compound TCP with Random Losses," in *Proceedings of the 8th International IFIP-TC 6 Networking Conference*, ser. NETWORKING '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 482–494.
- [38] S. Poojary and V. Sharma, "Approximate theoretical models for TCP connections using different high speed congestion control algorithms in a multihop network," in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, Oct 2013, pp. 559–566.
- [39] D. Ghosh, K. Jagannathan, and G. Raina, "Right buffer sizing matters: Stability, queuing delay and traffic burstiness in compound TCP," in *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, Sept 2014, pp. 1132–1139.
- [40] S. Chavan and G. Raina, "Performance of TCP with a Proportional Integral Enhanced (PIE) queue management policy," in *Control and Decision Conference (CCDC), 2015 27th Chinese*, May 2015, pp. 1013–1018.
- [41] S. Manjunath and G. Raina, "Analyses of compound TCP with Random Early Detection (RED) queue management," in *Control and Decision Conference (CCDC), 2015 27th Chinese*, May 2015, pp. 5334–5339.
- [42] S. Ha and S. Hemminger, "TCP CUBIC source code, ver 2.3," http://lxr.free-electrons.com/source/net/ipv4/tcp_cubic.c?v=4.2, Last accessed: 08 Dec 2015.
- [43] S. Asmussen, *Applied probability and queues*, 2nd ed., ser. Applications of Mathematics (New York). New York: Springer-Verlag, 2003, vol. 51, stochastic Modelling and Applied Probability.
- [44] D. Bertsekas and R. Gallager, *Data Networks – 2nd Edition*. Prentice-Hall, Inc., 1992.
- [45] C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," *Mathematics of computation*, pp. 577–593, 1965.
- [46] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.
- [47] A. Chaitanya, U. Mukherji, and V. Sharma, "Power allocation for interference channels," in *Communications (NCC), 2013 National Conference on*. IEEE, 2013.
- [48] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 8, pp. 1439–1451, Aug 2006.

- [49] D. X. Wei and P. Cao, “NS-2 TCP-Linux: an NS-2 TCP implementation with congestion control algorithms from Linux,” in *WNS2 '06: Proceeding from the 2006 workshop on ns-2: the IP network simulator*. New York, NY, USA: ACM Press, 2006.